Applied Probability Trust (20 June 2019)

INSENSITIVITY OF THE MEAN-FIELD LIMIT OF LOSS SYSTEMS UNDER SQ(d) ROUTING

THIRUPATHAIAH VASANTAM,* *University of Waterloo*

ARPAN MUKHOPADHYAY,** *University of Warwick*

RAVI R. MAZUMDAR,* *University of Waterloo*

Abstract

In this paper, we study a large multi-server loss model under the SQ(d) routing scheme when the service time distributions are general with finite mean. Previous works have addressed the exponential service time case when the number of servers goes to infinity giving rise to a mean-field model. The fixed-point of the limiting mean-field equations (MFEs) was seen to be insensitive to the service time distribution in simulations but no proof was available. While insensitivity is well known for loss systems models, even with state-dependent inputs, belong to the class of linear Markov models. In the context of SQ(d) routing, the resulting model belongs to the class of nonlinear Markov processes (processes whose generator itself depends on the distribution) for which traditional arguments do not directly apply. Showing insensitivity to the general service time distributions has thus remained an open problem. Obtaining the MFEs in this case poses a challenge due to the resulting Markov description of the system being in positive orthant as opposed to a finite chain in the exponential case. In this paper, we first obtain the MFEs and then show that the MFEs have a unique fixed point that coincides with the fixed point in the exponential case thus establishing insensitivity. The approach is via a measure-valued Markov process representation and the martingale problem to establish the mean-field limit.

Keywords: Erlang loss models; SQ(d); Mean-field; Fixed-point; Insensitivity

2010 Mathematics Subject Classification: Primary 60K35

Secondary 60F10;60J10;62F15

* Postal address: Department of Electrical and Computer Engineering, 200 University Ave W, Waterloo, ON N2L 3G1

** Postal address: Department of Computer Science, University of Warwick, Coventry, CV4 7AL

1. Introduction

We consider a multi-server loss system consisting of a large number N of parallel servers to which jobs arrive according to a Poisson process with rate $N\lambda$ and the service times are generally distributed with finite mean. Each server has capacity to serve up to C jobs simultaneously, and there is no waiting room. A central job dispatcher routes an incoming job to one of the servers where the processing of the job at unit rate begins immediately if the number of jobs that are already in progress is less than C otherwise, the job gets blocked or discarded. The job length is assumed to be random from a general distribution with finite mean. These models appear in practice in cloud computing systems such as Microsoft's Azure [30] and Amazon EC2 [2].

The motivation behind considering such models is that due to a tremendous growth in the trend to externalize storage and computing resources, cloud computing systems maintain a large number of servers to provide service to the incoming jobs. In these systems, the job requests are mapped into virtual machines (VMs) that request resources such as processor power, I/O bandwidth, disk etc. from a server that is picked from a large set of available servers. When a job arrives, the incoming request is routed to one of the servers where it is accepted for the service if the requested amount of resources are available, otherwise it is blocked or discarded. The resources allocated to a job will be released once the service of a job ends. In order to provide good quality of service, the service provider in cloud computing systems uses a routing policy at the job dispatcher that balances loads on servers that minimizes the average blocking probability or the probability that a request cannot be accommodated. Since the job requests arrive randomly and their durations are random too, the way this is achieved is to route arrivals to servers that are least loaded or have the smallest number of jobs. This is referred to as the join-the-shortest-queue (JSQ) policy and it requires knowledge of the occupancies of all the servers. Large cloud computing systems have thousands of servers and the individual server occupancies will need to be maintained at the dispatcher. However, this is not necessary as the randomized sampling of just a few servers has been shown to perform almost as well as complete sampling [31, 32, 45] for models of interest. This policy is referred to as the SQ(d) policy, short for the power-of-d routing policy, that routes an incoming request to the shortest of d uniformly sampled servers.

The SQ(d) scheme was first introduced in [45] for multi-server server systems with FCFS

service discipline for the case of $d = 2$ and exponential service times. When the number of servers N is finite, analyzing the SQ(d) routing policy is a difficult task due to dependence amongst the servers introduced by the SQ(d) policy. However, when $N \to \infty$, they obtained a tractable way of characterizing the stationary distributions that are accurate when the number N is large [45]. Their results were then extended for the case of $d > 2$ in [31] where it was argued that the case $d = 2$ provides most of the gains and hence the term 'The power-of-2' came to be used.

Loss models similar to the one considered here were analyzed in [35, 36, 46] under the assumption of exponential service time distributions for the SQ(d) routing policy. They also considered the more general heterogeneous case with an appropriate modification of the SQ(d) policy to account for server and job heterogeneity. It was shown in [36] that the SQ(d) routing scheme yields almost optimal blocking performance in that the average blocking is very close to the theoretical lower bound on the minimum average blocking achievable by any work conserving policy. In simulations the stationary occupancy distributions were observed to be insensitive to the service time distribution.

In the case of exponential service times, the results shown actually imply that the following interchange holds. Let $\mathbf{x}^N(t) = (x_l^N(t), l \geq 0)$ where $x_l^N(t)$ denotes the fraction of servers with at least l jobs. Then

$$\lim_{N\to\infty} \lim_{t\to\infty} \mathbf{x}^N(t) = \lim_{t\to\infty} \lim_{N\to\infty} \mathbf{x}^N(t). \tag{1}$$

Equation (1) provides the equivalence between the stationary distribution of the limiting system given by the left hand side and the globally stable fixed-point or equilibrium of the mean-field given by the the right hand side under the SQ(d) routing policy. The key is that the mean field equation is a deterministic differential equation that is easier to study. Moreover, this property can be used to show that in the limit, the individual systems are statistically independent.

In most applications, the service time distributions are not exponential. For example, the service times follow log-normal distributions in call centers [8], and Gamma distributions in automatic teller machines (ATMs) [26] etc. The focus of this paper is to consider this scenario and develop a mean-field model and characterize the properties of its fixed point.

For general service times case, a Markovian modeling of the system requires us to track the age or residual service time of each job that is in progress in the system. Therefore the

underlying space on which the Markov process lies is not discrete and hence the classical Markov chain techniques cannot be used. This makes establishing the mean-field limit and characterizing the properties of its equilibrium behavior for general service times a challenging task.

It is well known that the stationary distributions of single server loss systems even with state-dependent Poisson arrival rates are insensitive to the service time distribution, i.e., they only depend on the mean of the service times [9]. Hence, it is important to investigate whether the insensitivity property carries over to the systems with randomized routing such as the SQ(d) routing policy. When N is finite, randomized strategies result in the individual servers being coupled. It can be shown as in [5, 6] that when N is finite, the system is not insensitive since the SQ(d) policy does not satisfy the necessary condition of state-dependent arrival rates to be balanced. Insensitivity of the fixed or equilibrium point was observed for the limiting case ($i.e.$ when $N \to \infty$) via simulations in [35, 46] but no proofs were provided. One of the main objectives in this paper is to answer this question.

A mean-field analysis for processor sharing (PS) queues with the SQ(d) routing has been done in [33, 34] in the exponential service time distribution case. In [7], randomized routing schemes for queueing systems with general service time distributions when service disciplines are FCFS, PS, and LIFO were studied. The steady-state results were characterized by assuming the asymptotic independence of servers in the system. However the mean-field limit and its fixed-point were not studied in any detail.

In [24] mean-field techniques were used to study closed queueing networks with M customers and N queues with FCFS service discipline in which an exiting customer from a queue joins another queue chosen with probability $\frac{1}{N}$ from N queues. The mean-field was established for the regime when $\lim_{M,N\to\infty} \frac{M}{N} \to \alpha$. However, the equilibrium behavior of the system was not studied. Recently, the SQ(d) setting in a system of N FCFS servers where jobs arrive according to a time-inhomogeneous Poisson process and general i.i.d service times was studied in [1]. They obtained the mean-field for the case of general service time distributions for all finite intervals of time. However, the steady-state analysis was not investigated.

Multi-server loss models with randomized routing schemes were first studied in [42, 43] when job lengths are exponentially distributed using a formal mean-field approach. However, the existence and uniqueness of the fixed-point of the mean-field were not shown. In [35, 36, 46], the existence and uniqueness of the fixed-point of the mean-field for homogeneous loss

model of [42] was addressed. In [46], the existence and uniqueness was established under the asymptotic independence of servers ansatz while [35] showed the asymptotic independence (or propagation of chaos) and that the interchange of limits (1) holds. Propagation of chaos on path space had been earlier studied by [16, 17] in the context of alternate routing in circuit-switched networks.

Mean-field analysis and the fluid analysis of queues are closely related, the former usually in the space of measures and the latter on the sample paths. The fluid limit analysis of FCFS and Processor Sharing queues with general service time distributions has been studied using a measure-valued processes approach developed by Dawson [10] in [12, 18, 19, 25, 47]. In this paper, we use the ages of jobs to construct a measure-valued Markov process that models the system dynamics and we establish the mean-field limit of the empirical measure-valued process as in [10, 12]. Our approach is similar to [14] where the FCFS model is studied with exponential distributions under the SQ(d) policy. In the exponential case the set of server states is the space of non-negative integers $\mathbb{Z}_+$. In [14] the law of large numbers on path space is established by studying the limit of the sequence of empirical measures with samples in $\mathcal{M}_1(\mathcal{D}_{\mathbb{Z}_+}([0,\infty)))$ where $\mathcal{D}_{\mathbb{Z}_+}([0,\infty))$ is the space of right continuous functions with left limits in $\mathbb{Z}_+$ and $\mathcal{M}_1(\mathcal{D}_{\mathbb{Z}_+}([0,\infty)))$ is the space of probability measures on $\mathcal{D}_{\mathbb{Z}_+}([0,\infty))$. More recently, in [15] a functional central limit theorem (CLT) is derived for the FCFS model showing that under the CLT scaling the limiting process is a stable Ornstein-Uhlenbeck process and the exchange of limits holds for this regime.

In this paper, we obtain the mean-field for the SQ(d) routing in loss systems and we characterize the fixed-point or equilibrium of the mean-field equations. Unlike the exponential case, the MFEs are now partial differential equations. In particular, we show that the fixed-point is unique and moreover coincides with the fixed point of the MFEs in the exponential case. This establishes the insensitivity of the fixed point.

The rest of the paper is organized as follows: Section 2 describes the system model and the SQ(d) policy. In Section 3, we introduce the notation used in the paper. In Section 4, we derive a measure-valued representation for the state of the system. The main results of the paper are given in Section 5. We then establish the mean-field limit in Section 6. In Section 7, we prove the main result on the uniqueness of the fixed point of the MFEs and show that the fixed point is insensitive to the distribution, $i.e.$, it depends only on the mean service time. In Section 8 we provide numerical results that suggest the global asymptotic stability of the fixed-point of

the MFEs and hence the relation given in (1) indeed holds. Section 9 concludes the paper with some remarks and generalizations. Proofs of supplementary technical results are provided in the Appendices.

2. System model and the routing policy

We consider a system consisting of a large number N of parallel servers. Jobs arrive according to a Poisson process with rate $N\lambda$ and the job lengths are assumed to be $i.\,i.\,d.$ from a general distribution $G(\cdot)$ defined on $\mathbb{R}_+$. A central job dispatcher routes an incoming job to a server according to the SQ(d) policy defined below. We assume that each server has capacity to process up to a number C of jobs simultaneously and each job is processed at unit rate. At any time t, if a server is currently serving i jobs, then we say that the server has occupancy i and vacancy $C - i$ at time t. If an incoming job is routed to a server with occupancy C, then the job is blocked or discarded, otherwise the processing of the job begins immediately and it is processed at unit rate.

Definition 2.1. *SQ(d) or Power-of-d routing: An incoming job is routed to the server with the minimum occupancy among d servers that are selected randomly with replacement. Ties among servers are broken by choosing a server uniformly at random. The randomly chosen d servers are referred to as the potential destination servers and the server to which a job is routed is called the destination server.*

In the Definition 2.1, we assume sampling with replacement because of notational convenience and it is easy to show that the asymptotic results that are of interest in the paper are not affected whether we sample with or without replacement.

We assume that the service times have finite mean $\frac{1}{\mu}$ and the service time distribution denoted by $G(\cdot)$ on $[0,\infty)$ possesses a continuous density denoted by $g(\cdot)$. We make an assumption that $\overline{G}(\cdot)$ is supported on $[0,\infty)$ where $\overline{G}(.)$ denotes the complementary distribution. The hazard rate function of $G(\cdot)$ is defined as $\beta(x) = \frac{g(x)}{\overline{G}(x)} = \frac{g(x)}{1-G(x)}$ for $x \in [0,\infty)$. The hazard rate function β indicates the instantaneous rate at which the service of a job ends. More precisely, a job with age y (where y denotes the time since its arrival) at time t exits the server in the interval $[t, t+dt)$ with probability $\beta(y)dt$.

Assumption 2.1. *The hazard rate function β satisfies $\beta \in \mathcal{C}_b(\mathbb{R}_+)$ where $\mathcal{C}_b(\mathbb{R}_+)$ denotes the*

space of continuous bounded functions on nonnegative real line $\mathbb{R}_+$

Remark 2.1. The Assumption 2.1 is true for several classes of distributions such as Phase-Type distributions, Gamma distributions, Log-Normal distributions, and any Pareto distribution with finite mean.

3. Notation and terminology

We first introduce the notation which is used throughout the paper. Let $\mathbb{Z}$, $\mathbb{R}$ be the set of integers and real numbers, respectively. Further, let $\mathbb{Z}_+$, $\mathbb{R}_+$ be the set of nonnegative integers and nonnegative real numbers, respectively.

Function and measure spaces.

For any given metric space $\mathcal{E}$, let $\mathcal{K}_b(\mathcal{E}), \mathcal{C}_b(\mathcal{E}), \mathcal{C}_s(\mathcal{E})$ be the space of bounded measurable real valued functions, the space of bounded continuous real valued functions, and the space of continuous real valued functions with compact support, defined on $\mathcal{E}$, respectively. Furthermore, let $\mathcal{C}^1(\mathcal{E})$ be the space of once continuously differentiable real valued functions defined on $\mathcal{E}$ and let the subspace of functions in $\mathcal{C}^1(\mathcal{E})$ which have compact support be denoted by $\mathcal{C}^1_s(\mathcal{E})$. The space of bounded functions in $\mathcal{C}^1(\mathcal{E})$ whose first derivatives are also bounded is denoted by $\mathcal{C}^1_b(\mathcal{E})$. For any function $f \in \mathcal{K}_b(\mathcal{E})$, $h \in \mathcal{C}^1(\mathcal{E})$, we define

$$\|f\| = \sup_{x\in\mathcal{E}} |f(x)|\,, \quad \|h\|_1 = \|h\| + \|h'\|$$

where h' denotes the derivative of h. The space $\mathcal{C}_b(\mathcal{E})$ is equipped with the uniform topology, $i.e.$, we say that a sequence of functions $(f_n \in \mathcal{C}_b(\mathcal{E}), n \geq 1)$ converges to a function $f \in \mathcal{C}_b(\mathcal{E})$ if $\|f_n - f\| \to 0$ as $n \to \infty$. The space $\mathcal{C}^1(\mathcal{E})$ is equipped with the topology induced by the norm $\|\cdot\|_1$.

For a given metric space $\mathcal{E}$, let the Borel σ-algebra be denoted by $\mathcal{B}(\mathcal{E})$. Let the space of finite non-negative measures on $\mathcal{E}$ be denoted by $\mathcal{M}_F(\mathcal{E})$. We use the notation $\nu(B)$ and $\nu(\{y\})$ to denote the measure of a Borel set $B \in \mathcal{B}(\mathcal{E})$ and an element $y \in \mathcal{E}$ with respect to the measure $\nu \in \mathcal{M}_F(\mathcal{E})$, respectively. The space of probability measures is denoted by $\mathcal{M}_1(\mathcal{E})$. Also, let $\mathcal{M}^N_1(\mathcal{E}) \subset \mathcal{M}_1(\mathcal{E})$ be the subspace of probability measures defined as $\mathcal{M}^N_1(\mathcal{E}) = \{\nu \in \mathcal{M}_1(\mathcal{E}) : N\,\nu(B) \in \mathbb{Z}_+, \forall B \in \mathcal{B}(\mathcal{E})\}$. For any $\phi \in \mathcal{K}_b(\mathcal{E})$, $\nu \in \mathcal{M}_F(\mathcal{E})$,

we define

$$\langle \nu, \phi \rangle = \int_{\mathcal{E}} \phi(y)\nu(dy).$$

The space of measures $\mathcal{M}_F(\mathcal{E})$ is equipped with the weak topology induced by the weak convergence of measures.

The age of an active job is the time elapsed since its arrival. To model the dynamics of an Erlang loss system with capacity C for each server by a Markov process, we define the state of each server as $(n, a_1, a_2, \cdots, a_n)$ where n denotes the number of jobs that are in progress at the server and a_i denotes the age of the i^{th} job in progress. We now define a space $\mathcal{U}$ that is used in earlier works to study queuing models with general service time distributions by using the classical supplement variable method [29, 41] such that it contains all the possible server states as elements. The space $\mathcal{U}$ is defined as

$$\mathcal{U} = \cup_{n=0}^{C} \mathcal{U}_n,$$

where $\mathcal{U}_0 = \{0\}$ and an element in $\mathcal{U}_n$ for $n \geq 1$ is of the form $(n, a_1, \ldots, a_n)$ where $1 \leq n \leq C$ and $a_i \in \mathbb{R}_+$. We specify that the state of a server that has n jobs belongs to the space $\mathcal{U}_n$. Here, one might omit the variable n and consider just $(a_1, \cdots, a_n)$ to represent a server state, but such representation does not account for idle servers while (0) is the state of idle servers in our representation. Furthermore, the variable n, directly gives us information about the number of progressing jobs at a server which changes upon every arrival and departure. Hence, it is convenient to work with the server state representation that has a variable that denotes the number of progressing jobs at a server.

It is also possible to define an element in $\mathcal{U}_n$ by $(n, a_1, \cdots, a_n, 0, \cdots, 0)$ of size $C + 1$. This allows us to have constant size of $C + 1$ for an element in $\mathcal{U}$. Note that the zeros in the state $(n, a_1, \cdots, a_n, 0, \cdots, 0)$ act as dummy variables as there are only n jobs. Hence, to make it simple, we consider an element in $\mathcal{U}_n$ is of the form $(n, a_1, \cdots, a_n)$ with size $n + 1$. Without loss of generality, we refer to an element in the set $\mathcal{U}$ by $\mathbf{u}$ and an element in the set $\mathcal{U}_n$ by $\mathbf{u}_n$. Note that we have $\mathbf{u}_0 = 0$. For $\mathbf{y}_n = (n, y_1, \ldots, y_n), \mathbf{z}_m = (m, z_1, \ldots, z_m)$, we define the metric $d_{\mathcal{U}}(\mathbf{y}_n, \mathbf{z}_m)$ as

$$d_{\mathcal{U}}(\mathbf{y}_n, \mathbf{z}_m) = \begin{cases} \sum_{i=1}^{n} |y_i - z_i| & \text{if } n = m, \\ \infty & \text{otherwise.} \end{cases}$$

For $(n, u_1, \ldots, u_n) \in \mathcal{U}_n$ and $y \geq 0$, we use the following notation

$$\begin{aligned}
\mathbf{u}_n &= (n, u_1, \cdots, u_n), \\
\mathbf{u}_n^{-j} &= (n-1, u_1, \cdots, u_{j-1}, u_{j+1}, \cdots, u_n), \\
(\mathbf{u}_n^{j}; y) &= (n+1, u_1, \cdots, u_{j-1}, y, u_j, \cdots, u_n), \\
(\mathbf{u}_n^{-j}; y) &= (n, u_1, \cdots, u_{j-1}, y, u_{j+1}, \cdots, u_n).
\end{aligned}$$

For any Borel set $B \in \mathcal{B}(\mathcal{U})$, let $I_{\{B\}}$ be the indicator function of B. Let the function $\mathbf{1}$ be defined such that for all $\mathbf{u} \in \mathcal{U}$, we have $\mathbf{1}(\mathbf{u}) = 1$.

A function $f : \mathcal{U} \mapsto \mathbb{R}$ is said to be differentiable if for every $n \geq 1$, the function $\frac{\partial f(\mathbf{u}_n)}{\partial u_i}$ exists for all $1 \leq i \leq n$ at every $\mathbf{u}_n \in \mathcal{U}_n$. As a result, the function $I_{\{\mathcal{U}_n\}}, n \geq 1$ is differentiable. For a differentiable function $f : \mathcal{U} \mapsto \mathbb{R}$, we have

$$\|f'\| = \max_{n \geq 1} \left(\sup_{\mathbf{u}_n \in \mathcal{U}_n} \left(\max_{1 \leq i \leq n} \left| \frac{\partial f(\mathbf{u}_n)}{\partial u_i} \right| \right) \right).$$

Further, for a differentiable function $f : \mathcal{U} \mapsto \mathbb{R}$, let the function $\nabla_1 f$ be defined as

$$\nabla_1 f(n, u_1, \cdots, u_n) = \nabla f \cdot \mathbf{1} = \sum_{i=1}^{n} \frac{\partial f(\mathbf{u}_n)}{\partial u_i}. \tag{2}$$

A measure $\nu \in \mathcal{M}_F(\mathcal{U})$ when it is restricted to $\mathcal{U}_0$ is a Dirac measure at $\{0\}$ satisfying $\nu(\mathcal{U}_0) = \nu(\{0\})$. We say that a measure ν is absolutely continuous with respect to Lebesgue measure if $\nu(\{\mathbf{x}_n\}) = 0$ at every $\mathbf{x}_n \in \mathcal{U}_n$ for all $n \geq 1$. For any Borel measurable function f that is defined on $\mathcal{U}$, we define

$$\langle \nu, f \rangle = f(0)\nu(\{0\}) + \sum_{n=1}^{C} \int_{\mathcal{U}_n} f(\mathbf{z}_n)\nu(d\mathbf{z}_n).$$

We now define the function $\mathcal{I} : \mathcal{U} \mapsto \mathbb{R}$ as follows:

$$\mathcal{I}(\mathbf{x}_n) = \begin{cases} \sum_{i=1}^{n} x_i & \text{if } n \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

For $b \geq 0$, let $\tau_b^+ : \mathcal{U} \mapsto \mathcal{U}$ be the transition operator defined as

$$\tau_b^+(\mathbf{x}_n) = \begin{cases} (n, x_1 + b, \cdots, x_n + b) & n \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for any $b \geq 0$ and $f \in \mathcal{K}_b(\mathcal{U})$, let the mapping $\tau_b : \mathcal{K}_b(\mathcal{U}) \mapsto \mathcal{K}_b(\mathcal{U})$ be defined as $\tau_y f(\mathbf{u}) = f(\tau_y^+ \mathbf{u})$. Also, for $b \geq 0$, let the measure $\tau_b \nu \in \mathcal{M}_F(\mathcal{U})$ be defined such that for any Borel set $B \in \mathcal{B}(\mathcal{U})$, $\tau_b \nu(B) = \nu(\tau_b^+(B))$. For $\nu \in \mathcal{M}_F(\mathcal{U})$, the measure $\tau_b \nu \in \mathcal{M}_F(\mathcal{U})$ satisfies $\langle \tau_b \nu, f \rangle = \langle \nu, \tau_b f \rangle$ for all $f \in \mathcal{K}_b(\mathcal{U})$ and the existence of the unique measure $\tau_b \nu$ follows from the Riesz-Markov-Kakutani theorem [40, Theorem 2.14].

Measure valued stochastic processes.

For a Polish space $\mathcal{H}$ and a nonnegative real number $T < \infty$, let the càdlàg functions, also referred to as RCLL (right continuous with left limits) functions, that are defined on $[0, T]$ and $[0, \infty)$ with values in $\mathcal{H}$ be denoted by $\mathcal{D}_{\mathcal{H}}([0, T])$ and $\mathcal{D}_{\mathcal{H}}([0, \infty))$ respectively. Similarly, let the space of continuous functions that take values in $\mathcal{H}$ defined on $[0, T]$ (resp. $[0, \infty)$) be denoted by $\mathcal{C}_{\mathcal{H}}([0, T])$ and $\mathcal{C}_{\mathcal{H}}([0, \infty))$, respectively. The spaces $\mathcal{D}_{\mathcal{H}}([0, T])$ and $\mathcal{D}_{\mathcal{H}}([0, \infty))$ are equipped with the Skorohod J_1-topology and hence, are Polish spaces. Let the covariation of two local martingales $(M_t^1, t \geq 0)$ and $(M_t^2, t \geq 0)$ in $\mathcal{D}_{\mathbb{R}}([0, T])$ be denoted by $(< M^1, M^2 >_t, t \geq 0)$ and the quadratic variation of $(M_t^1, t \geq 0)$ be denoted by $(< M^1 >_t, t \geq 0) = (< M^1, M^1 >_t, t \geq 0)$.

In our analysis, we study $\mathcal{H}-$valued stochastic processes where $\mathcal{H} = \mathcal{M}_F(\mathcal{U})$. The considered stochastic processes are random elements defined on $(\Omega, \mathbb{F}, \mathbb{P})$ with sample paths in $\mathcal{D}_{\mathcal{H}}([0, \infty))$, and are equipped with the Borel $\sigma-$algebra generated by the open sets under the Skorohod J_1- topology [4]. We say that a sequence of stochastic processes $\{X_n\}_{n \geq 1}$ where X_n is defined on $(\Omega_n, \mathbb{F}_n, \mathbb{P}_n)$ with sample paths lying in $\mathcal{D}_{\mathcal{H}}([0, \infty))$ converges in distribution to a stochastic process X defined on $(\Omega, \mathbb{F}, \mathbb{P})$ with sample paths lying in $\mathcal{D}_{\mathcal{H}}([0, \infty))$, if for every bounded, continuous, real valued functional $F : \mathcal{D}_{\mathcal{H}}([0, \infty) \to \mathbb{R}$, we have $\lim_{n \to \infty} \mathbb{E}_n(F(X_n)) = \mathbb{E}(F(X))$ where the expectation operators $\mathbb{E}_n, \mathbb{E}$ are defined with respect to $\mathbb{P}_n, \mathbb{P}$, respectively. We denote the convergence of $\{X_n\}_{n \geq 1}$ in distribution to X by $X_n \Rightarrow X$.

4. State descriptor and system dynamics

We index a sequence of systems by N that denotes the total numbers of servers. Incoming jobs arrive according to a Poisson process with rate $N\lambda$ and the job lengths are i.i.d. from a common distribution $G(\cdot)$ defined on $\mathbb{R}_+$. The state of a server is written as

$\mathbf{a}_n = (n, a_1, \cdots, a_n) \in \mathcal{U}$ when there are n progressing jobs and i^{th} job has age a_i for $1 \leq i \leq n$. A server with state say $\mathbf{a}_n$ can be viewed as a particle with the given state. Therefore the system evolution can be considered as the evolution of a system with N particles where the interactions between particles takes place while routing an arrival according to the SQ(d) routing policy.

The age of a job that is in service at a server increases linearly with time at unit rate until its service is completed. We next describe the possible state of a server at time $t + h$ ($h > 0$) given that it has state $\mathbf{a}_n$ at time t. We assume that when h is small enough, in the interval $[t, t+h)$, the probability of having multiple events of arrivals or departures is negligible. In the interval $[t, t+h)$, if there is no arrival or departure at the given server, then the server state will be equal to $\tau_h^+(\mathbf{a}_n)$ at time $t + h$. On the other hand, if i^{th} job expires in the interval $[t, t + h)$, then the server state will be equal to $\tau_h^+(\mathbf{a}_n^{-i})$ at time $t+h$. Considering arrivals, suppose there is an arrival into the server at time $t + r$ ($0 \leq r < h$), then the arriving job chooses its position uniformly at random out of $n + 1$ possible positions and suppose it chooses j^{th} position, then the server state will be equal to $((\tau_h^+(\mathbf{a}_n))^j; h - r)$ at time $t + h$.

Let $S^N_{(i,t)} \in \mathcal{U}$ be the random variable that indicates the state of the server i at time t. Although, one can think of considering $(S^N_{(1,t)}, \cdots, S^N_{(N,t)})$ to denote the system state at time t which is a Markovian representation of the system, the dimension of this state space increases with N as $N \to \infty$ which is inconvenient to work with since our focus of interest is to study the asymptotic behavior of the system as $N \to \infty$. Hence, we consider an alternative simple system state representation that can be used to describe the system evolution as the evolution of a Markov process. Note that the system is symmetric with respect to the servers as they are identical and the server identities do not play any role in the evolution with time. Therefore to model the system evolution by a Markov process, we will show that it is enough to just keep track of the number of servers that lie in each state $\mathbf{u} \in \mathcal{U}$ in order to establish the mean-field limit. Measure-valued Markov processes have also been used to study other interacting particle systems as in [20, 28, 37] where each particle state $\mathbf{x} \in \mathbb{R}^n$, $n > 1$ is viewed as a measure-valued Markov process. Following these works, we consider the following system state descriptor.

Definition 4.1. *At time t, the state descriptor of the system with index N is a random measure*

given by

$$\eta_t^N = \sum_{i=1}^{N} \delta_{S^N_{(i,t)}}. \tag{3}$$

The interpretation of η_t^N is that for any measurable function f defined on $\mathcal{U}$, we have

$$\langle \eta_t^N, f \rangle = \sum_{i=1}^{N} f(S^N_{(i,t)}).$$

At time t, conditioned on server states say $S^{(N)}_{(i,t)} = s_{(i,t)}$, the system state can be represented by a measure ν defined as

$$\nu = \sum_{i=1}^{N} \delta_{s_{(i,t)}}. \tag{4}$$

For $\eta_t^N = \nu$, an element $\mathbf{y} \in \mathcal{U}$ is an atom of ν if there exists at least one server with the state $\mathbf{y}$ at time t. The mass of an atom of ν denotes the number of servers lying at that atom at time t. As a result, since the number of interacting particles in the system is equal to N, the measure ν defined on $\mathcal{U}$ contains a finite number of atoms which is bounded by N. If all the servers have different states then the number of atoms is equal to N, otherwise the number of atoms is less than N. Let V_t be the number of atoms at time t and let the i^{th} atom be denoted by $\mathbf{v}_t^{(i)}$. Further, let the mass of the atom $\mathbf{v}_t^{(i)}$ be denoted by $a_t^{(i)}$. Here, $a_t^{(i)}$ denotes the number of servers that lie in the state $\mathbf{v}_t^{(i)}$ at time t and $a_t^{(i)} \geq 1$. Hence, for time t, from (4), we can also write ν as

$$\nu = \sum_{i=1}^{V_t} a_t^{(i)} \delta_{\mathbf{v}_t^{(i)}}. \tag{5}$$

For any Borel set $B \in \mathcal{B}(\mathcal{U})$, the number of servers with ages lying in the set B is equal to $\eta_t^N(B) = \nu(B) = \langle \nu, I_{\{B\}} \rangle$. We now define the measure of an element $\mathbf{y}_n = (n, y_1, \cdots, y_n)$ as below. Let $B^\epsilon(\mathbf{y}_n) = \{(n, r_1, \cdots, r_n) : y_i \leq r_i < y_i + \epsilon, 1 \leq i \leq n\}$. Then as in [18], we define

$$\nu(\{\mathbf{y}_n\}) = \lim_{\epsilon \to 0} \nu(B^\epsilon(\mathbf{y}_n)). \tag{6}$$

Essentially, $\nu(\{\mathbf{y}_n\})$ indicates the number of servers with state $\mathbf{y}_n$ at time t and can be viewed as an occupation count. The notation $d\nu(\mathbf{y}_n)$ denotes the number of servers with state lying in the interval $[\mathbf{y}_n, \mathbf{y}_n + d\mathbf{y}_n)$, where $d\mathbf{y}_n = (dy_1, \cdots, dy_n)$ and $\mathbf{y}_n + d\mathbf{y}_n$ is the vector addition of $\mathbf{y}_n$ and $d\mathbf{y}_n$. If there is no server lying in the state $\mathbf{y}_n$ at time t, then $\nu(\{\mathbf{y}_n\}) = 0$, otherwise $\mathbf{y}_n$ is an atom with mass $\nu(\{\mathbf{y}_n\})$. The number of servers that have n progressing jobs at time t is given by $\nu(\mathcal{U}_n) = \langle \nu, I_{\{\mathcal{U}_n\}} \rangle$.

We now obtain the probability that the destination server of an arrival lies in a particular state.

Lemma 4.1. *At time t, given that the system state is ν, i.e., $\eta_t^N = \nu$, under the SQ(d) routing policy, the probability that the destination server of an arrival at time t lies in the state $\mathbf{z}_n = (n, z_1, \cdots, z_n)$ where $\mathbf{z}_n$ is an atom of ν is given by*

$$p_r(\nu : \mathbf{z}_n) = \frac{\nu(\{\mathbf{z}_n\})}{N} \frac{(\overline{R}_n(\frac{\nu}{N})^d - \overline{R}_{n+1}(\frac{\nu}{N})^d)}{(\overline{R}_n(\frac{\nu}{N}) - \overline{R}_{n+1}(\frac{\nu}{N}))}, \tag{7}$$

where $\overline{R}_n(\frac{\nu}{N}) = \sum_{j:j=n}^{C} \frac{\nu}{N}(\mathcal{U}_j)$ represents the fraction of servers with at least n jobs.

Proof. When a potential destination server is chosen uniformly at random from N servers, it will have state $(n, z_1, \ldots, z_n)$ with probability $\frac{\nu(\{(n,z_1,\ldots,z_n)\})}{N}$. Suppose out of the d potential destination servers, say j servers have occupancy n and the remaining $d-j$ servers have occupancy at least $n+1$. Further, out of the j $(j \geq 1)$ potential destination servers with occupancy n, assume r $(r \geq 1)$ servers lie in the state $\mathbf{z}_n$. Then the probability that the destination server is a server with state $\mathbf{z}_n$ is given by

$$\binom{d}{j}\binom{j}{r}\left(\frac{r}{j}\right)\left(\frac{\nu(\{\mathbf{z}_n\})}{N}\right)^r\left(\frac{\nu(\{\mathcal{U}_n\}) - \nu(\{\mathbf{z}_n\})}{N}\right)^{j-r}\left(\sum_{i:i=n+1}^{C}\frac{\nu(\mathcal{U}_i)}{N}\right)^{d-j}.$$

Finally, by summing over all the possible values of j $(j \geq 1)$ and r $(r \geq 1)$, we have

$$\begin{aligned}
&\sum_{j=1}^{d}\sum_{r=1}^{j}\binom{d}{j}\binom{j}{r}\left(\frac{r}{j}\right)\left(\frac{\nu(\{\mathbf{z}_n\})}{N}\right)^r\left(\frac{\nu(\{\mathcal{U}_n\}) - \nu(\{\mathbf{z}_n\})}{N}\right)^{j-r}\left(\sum_{i:i=n+1}^{C}\frac{\nu(\mathcal{U}_i)}{N}\right)^{d-j} \\
&\qquad = \sum_{j=1}^{d}\binom{d}{j}\frac{1}{j}\left(\sum_{i:i=n+1}^{C}\frac{\nu(\mathcal{U}_i)}{N}\right)^{d-j}\left(\frac{\nu(\{\mathcal{U}_n\})}{N}\right)^j \\
&\qquad\qquad \times\left[\sum_{r=1}^{j} r\binom{j}{r}\left(\frac{\frac{\nu(\{\mathbf{z}_n\})}{N}}{\frac{\nu(\{\mathcal{U}_n\})}{N}}\right)^r\left(\frac{\frac{\nu(\{\mathcal{U}_n\})}{N} - \frac{\nu(\{\mathbf{z}_n\})}{N}}{\frac{\nu(\{\mathcal{U}_n\})}{N}}\right)^{j-r}\right].
\end{aligned}$$

The term inside the square bracket in the above equation is the average of a binomial random variable and hence, it is equal to $j\left(\frac{\frac{\nu(\{\mathbf{z}_n\})}{N}}{\frac{\nu(\{\mathcal{U}_n\})}{N}}\right)$. As a result, the above expression simplifies to

$$p_r(\nu : \mathbf{z}_n) = \left(\frac{\frac{\nu(\{\mathbf{z}_n\})}{N}}{\frac{\nu(\{\mathcal{U}_n\})}{N}}\right)\sum_{j=1}^{d}\binom{d}{j}\left(\sum_{i:i=n+1}^{C}\frac{\nu(\mathcal{U}_i)}{N}\right)^{d-j}\left(\frac{\nu(\{\mathcal{U}_n\})}{N}\right)^j.$$

We can further write

$$p_r(\nu : \mathbf{z}_n) = \left(\frac{\frac{\nu(\{\mathbf{z}_n\})}{N}}{\frac{\nu(\{\mathcal{U}_n\})}{N}}\right)\left[\left(\sum_{j=0}^{d}\binom{d}{j}\left(\sum_{i:i=n+1}^{C}\frac{\nu(\mathcal{U}_i)}{N}\right)^{d-j}\left(\frac{\nu(\{\mathcal{U}_n\})}{N}\right)^j\right) - \left(\sum_{i:i=n+1}^{C}\frac{\nu(\mathcal{U}_i)}{N}\right)^d\right].$$

After simplifications, we get (7). □

Remark 4.1. We can also interpret the expression of the $p_r(\nu : \mathbf{z}_n)$ as follows: The probability that all the potential destination servers have occupancy at least n and there exists at least one potential destination server with occupancy n is equal to $\overline{R}_n(\frac{\nu}{N})^d - \overline{R}_{n+1}(\frac{\nu}{N})^d$. From the SQ($d$) policy, the probability that the destination has occupancy n is equal to $\overline{R}_n(\frac{\nu}{N})^d - \overline{R}_{n+1}(\frac{\nu}{N})^d$. From the list of the servers with occupancy n, the fraction of the servers with the state $\mathbf{z}_n$ is equal to $\frac{\left(\frac{\nu(\{\mathbf{z}_n\})}{N}\right)}{\left(\frac{\nu(\{\mathcal{U}_n\})}{N}\right)}$. Therefore the probability that the destination server lies in the state $\mathbf{z}_n$ is equal to $\frac{\left(\frac{\nu(\{\mathbf{z}_n\})}{N}\right)}{\left(\frac{\nu(\{\mathcal{U}_n\})}{N}\right)} \times (\overline{R}_n(\frac{\nu}{N})^d - \overline{R}_{n+1}(\frac{\nu}{N})^d)$.

For the case of exponential job length distributions, $\mathcal{U}_n = \{n\}$ and $\mathbf{z}_n = n$. Hence, $p_r(\nu : \mathbf{z}_n) = \overline{R}_n(\frac{\nu}{N})^d - \overline{R}_{n+1}(\frac{\nu}{N})^d$ coinciding with the analysis for the exponential case in [31, 35].

As it is clear from (7), the routing decision depends only on the number of servers lying in each possible server state. Hence, we get the evolution of the process $(\eta_t^N, t \geq 0)$ by tracking arrival events, routing decisions, and departure events.

5. Main results

Our aim is to study the limit as $N \to \infty$ of the empirical measure of the distribution of the servers. For this, we define a sequence of systems such that a system with index N has N servers that process the incoming jobs arriving according to a Poisson process with rate $N\lambda$, and all other system parameters remain the same for all N as given in the Section 2. The system consists of a central job dispatcher that routes an arrival to a server according to the SQ(d) policy. For given N, the process $(\eta_t^N, t \geq 0)$ defined in equation (3) describes the dynamics of the system with index N. The goal is to characterize the limit of the normalized process $(\overline{\eta}_t^N, t \geq 0)$ as $N \to \infty$ where

$$\overline{\eta}_t^N = \frac{\eta_t^N}{N}. \tag{8}$$

For a Borel set $B \in \mathcal{B}(\mathcal{U})$, $\overline{\eta}_t^N(B)$ is equal to the fraction of the servers with state lying in the set B at time t.

5.1. Summary of analysis

We now give a brief overview of the analysis in the paper.

The mean-field limit corresponds to $\lim_{N\to\infty} \overline{\eta}_t^N = \overline{\eta}_t, \quad t \geq 0$, that takes values in $\mathcal{C}_{\mathcal{M}_1(\mathcal{U})}([0,\infty))$ and is a deterministic measure-valued process satisfying a set of evolution equations referred to as the mean-field equations. We then obtain an alternative form of the evolution equations satisfied by the process $(\langle\overline{\eta}_t, \psi\rangle, t \geq 0)$ for $\psi \in \mathcal{C}_b(\mathcal{U})$. This is stated in Lemma 5.1. Using these equations, we show in Theorem 5.1 that there exists a unique solution to the mean-field equations for a given initial point.

We then show that the sequence of processes $\{(\overline{\eta}_t^N, t \geq 0)\}$ is tight. For this, we first study the Feller property of the Markov process $(\eta_t^N, t \geq 0)$ and obtain the expression of its semigroup operator in Appendix A. In Appendix D, we construct a martingale process by using the generator of the Markov process $(\overline{\eta}_t^N, t \geq 0)$ by employing the Dynkin's formula [11, Theorem 7.15]. We then show that the martingale process converges to the null process as $N \to \infty$. Using this we prove the tightness of the sequence of processes $\{(\overline{\eta}_t^N, t \geq 0)\}$. Furthermore, we show that any limit point of the normalized process $(\overline{\eta}_t^N, t \geq 0)$ coincides almost surely with the unique solution to the mean-field equations referred to as the mean-field limit. This is stated in Theorem 5.2.

Finally, we obtain a set of the partial differential equations satisfied by the mean-field limit. We then prove the uniqueness of the fixed-point and its insensitivity. This is stated in Theorem 5.3. The proofs of Theorem 5.2 and Theorem 5.3 are given in Section 6 and Section 7, respectively. The remaining proofs are given in the Appendix.

5.2. Transient regime:

In this section, we discuss the results on the transient regime. For given system parameters λ, C, d and the probability density function $g(\cdot)$ of the service time distributions, in Proposition 5.1 we state the mean-field equations. The dynamics of a mean-field solution $(\overline{\eta}_t, t \geq 0)$ are described by using a set of evolution equations of the real valued processes $(\langle\overline{\eta}_t, \phi\rangle, t \geq 0)$ for all $\phi \in \mathcal{C}_b^1(\mathcal{U})$, referred to as the mean-field equations.

Proposition 5.1. *Mean-field equations:*

For given system parameters $(\lambda, C, d, g(\cdot))$, the process $(\overline{\eta}_t, t \geq 0)$ satisfies:

1. *The mapping $t \mapsto \overline{\eta}_t$ is a continuous.*

2. For $\phi \in \mathcal{C}_b^1(\mathcal{U})$, the process $(\overline{\eta}_t, t \geq 0)$ satisfies

$$\begin{aligned}\langle \overline{\eta}_t, \phi \rangle = \langle \overline{\eta}_0, \phi \rangle + \int_{s=0}^{t} \langle \overline{\eta}_s, \nabla_1 \phi \rangle \, ds \\ - \int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n) \right) d\overline{\eta}_s(\mathbf{x}_n) \right) ds \\ + \int_{s=0}^{t} \left(\overline{\eta}_s(\{0\}) \lambda \Phi_0(\overline{\eta}_s) \left(\phi(1,0) - \phi(0) \right) + \sum_{n=1}^{C-1} \sum_{i=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \right. \\ \left. \times \lambda \Phi_n(\overline{\eta}_s) (\phi(\mathbf{x}_n^i; 0) - \phi(\mathbf{x}_n)) \, d\overline{\eta}_s(\mathbf{x}_n) \right) ds, \quad (9)\end{aligned}$$

where the index j is used to denote the position of the departing job when there are n progressing jobs and i denotes the position of the arriving job when there are already n progressing jobs at the server. Further, $\Phi_n(\overline{\eta}_s) = \frac{(\overline{R}_n(\overline{\eta}_s)^d - \overline{R}_{n+1}(\overline{\eta}_s)^d)}{(\overline{R}_n(\overline{\eta}_s) - \overline{R}_{n+1}(\overline{\eta}_s))}$ where $\overline{R}_j(\overline{\eta}_s) = \sum_{n:n=j}^{C} \overline{\eta}_s(\mathcal{U}_n)$.

In (9), the second term on the right hand side is due to the increase of the ages of the progressing jobs linearly with time at unit rate. The third and fourth terms on the right hand side of (9) are due to the departure and arrival of a job, respectively.

Remark 5.1. The t-continuity of the mapping $\overline{\eta}_t$ is equivalent to the continuity of the mapping $t \mapsto \langle \overline{\eta}_t, \phi \rangle$ for all $\phi \in \mathcal{C}_b^1(\mathcal{U})$ since $\mathcal{C}_b^1(\mathcal{U})$ is a separating class of $\mathcal{M}_1(\mathcal{U})$ [13, p. 111].

Although the mean-field equation (9) is defined for the class of functions $\phi \in \mathcal{C}_b^1(\mathcal{U})$, it is more useful to obtain an approximation of the process $(\langle \overline{\eta}_t^N, I_{\{B\}} \rangle, t \geq 0)$ for an open set $B \in \mathcal{B}(\mathcal{U})$. Therefore we need to obtain the evolution equations of the real valued process $(\langle \overline{\eta}_t, I_{\{B\}} \rangle, t \geq 0)$. In this direction, we first obtain the evolution equations of the real valued process $(\langle \overline{\eta}_t, \psi \rangle, t \geq 0)$ where $\psi \in \mathcal{C}_b(\mathcal{U})$. We then proceed to obtain the evolution equations of the process $(\langle \overline{\eta}_t, I_{\{B\}} \rangle, t \geq 0)$ where B is an open set with the help of the monotone convergence theorem since there exists a sequence of functions in $\mathcal{C}_b(\mathcal{U})$ that increase point wise to $I_{\{B\}}$.

Lemma 5.1. *A process $(\nu_t \in \mathcal{M}_1(\mathcal{U}), t \geq 0)$ with continuity of the mapping $t \mapsto \nu_t$ satisfies*

the mean-field equation (9) *iff it satisfies the following equation for all* $\phi \in \mathcal{C}_b(\mathcal{U})$*,*

$$\langle \nu_t, \phi \rangle = \langle \nu_0, \tau_t \phi \rangle + \int_{r=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\tau_{t-r}\phi(\mathbf{x}_n^{-j}) - \tau_{t-r}\phi(\mathbf{x}_n) \right) d\nu_r(\mathbf{x}_n) \right.$$
$$+ \left[\nu_r(\{0\}) \lambda \Phi_0(\nu_r) \left(\tau_{t-r}\phi(1,0) - \tau_{t-r}\phi(0) \right) \right.$$
$$\left. \left. + \sum_{n=1}^{C-1} \sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \lambda \Phi_n(\nu_r) (\tau_{t-r}\phi(\mathbf{x}_n^j; 0) - \tau_{t-r}\phi(\mathbf{x}_n)) \, d\nu_r(\mathbf{x}_n) \right] \right) dr. \quad (10)$$

The proof is given in Appendix B.

Using equation (10), we show that starting with an initial measure ν_0, for $t \geq 0$, there exists a unique measure $\nu_t \in \mathcal{M}_1(\mathcal{U})$ that satisfies equation (9).

For any finite measure ν defined on $\mathcal{U}$, the operator $\langle \nu, \phi \rangle$ is a continuous linear operator on the space of functions $\phi \in \mathcal{C}_b(\mathcal{U})$ and let

$$\|\nu\| = \sup_{\phi \in \mathcal{C}_b(\mathcal{U})} \frac{|\langle \nu, \phi \rangle|}{\|\phi\|}. \quad (11)$$

Theorem 5.1. *There exists a unique solution in* $\mathcal{C}_{\mathcal{M}_1(\mathcal{U})}([0,\infty))$ *to the mean-field equations. In particular, if* $(\nu_t^1, t \geq 0)$ *and* $(\nu_t^2, t \geq 0)$ *are two mean-field solutions starting at initial measures* $\nu_0^1 \in \mathcal{M}_1(\mathcal{U})$, $\nu_0^2 \in \mathcal{M}_1(\mathcal{U})$*, respectively, then*

$$\|\nu_t^1 - \nu_t^2\| \leq \|\nu_0^1 - \nu_0^2\| \, e^{(2C\|\beta\| + 8d^2\lambda)t}. \quad (12)$$

The proof is given in Appendix C.

We now show the convergence of the sequence of the processes $(\overline{\eta}_t^N, t \geq 0)$. For this, we first assume:

Assumption 5.1. *The sequence of the initial random measures* $\{\overline{\eta}_0^N\}$ *satisfy*

$$(\overline{\eta}_0^N, \langle \overline{\eta}_0^N, \mathcal{I} \rangle) \Rightarrow (\vartheta, \langle \vartheta, \mathcal{I} \rangle), \quad (13)$$

where $\vartheta \in \mathcal{M}_1(\mathcal{U})$ *is a probability measure that possesses a density (w.r.t. Lebesgue measure) and* $\langle \vartheta, \mathcal{I} \rangle < \infty$*.*

Theorem 5.2. *If the sequence of random measures* $\{\overline{\eta}_0^N\}$ *satisfies the Assumption 5.1, then* $(\overline{\eta}_t^N, t \geq 0) \Rightarrow (\overline{\eta}_t, t \geq 0)$*, where* $(\overline{\eta}_t, t \geq 0)$ *is the unique solution to the equation* (9) *with the initial point* ϑ*. The process* $(\overline{\eta}_t, t \geq 0)$ *is referred to as the mean-field limit.*

The proof is given in Section 6.

Remark 5.2. For any time t, a consequence of Theorem 5.2 is that as $N \to \infty$, any finite set of servers are independent of each other. Furthermore, as $N \to \infty$, $\overline{\eta}_t$ indicates the probability law of a server's state at time t and the arrival process to a server is a Poisson process with rate $\lambda\Phi_n(\overline{\eta}_t)$ when there are n ($n \geq 0$) progressing jobs. The proof follows from the same arguments as in the proof of the Proposition 2 of [35].

Lemma 5.2. *For any time t, the measure $\overline{\eta}_t$ has a density function $w.r.t.$ Lebesgue measure for almost all $\mathbf{u}_n \in \mathcal{U}_n$, $n \geq 1$.*

The proof is given in Appendix F.

For any subset $B \in \mathcal{B}(\mathcal{U})$, once $(\overline{\eta}_t^N, t \geq 0) \Rightarrow (\overline{\eta}_t, t \geq 0)$, since $\overline{\eta}_t$ is absolutely continuous $w.r.t.$ Lebesgue measure for every $t \geq 0$, the continuous mapping theorem implies that $(\langle\overline{\eta}_t^N, I_{\{B\}}\rangle, t \geq 0) \Rightarrow (\langle\overline{\eta}_t, I_{\{B\}}\rangle, t \geq 0)$. This shows that for large N, we can approximate $\langle\overline{\eta}_t^N, I_{\{B\}}\rangle$ by $\langle\overline{\eta}_t, I_{\{B\}}\rangle$.

5.3. Stationary regime:

We now discuss the stationary behavior of the mean field.

We first demonstrate an analogy between the MFEs of the considered multi-server Erlang loss system under the SQ(d) routing policy and the dynamics of an another single server Erlang loss system with state-dependent arrivals. We then exploit this analogy to prove the uniqueness of the fixed-point of the mean-field and its insensitivity. We first recall the dynamics of the probability measure of the server state of a single server Erlang loss system with capacity C, where jobs arrive according to a Poisson process with pre-specified state-dependent arrival rates.

Consider a single server system with capacity C where jobs arrive according to a Poisson process at rate α_n when there are n progressing jobs in the system. The service times are generally distributed as stated in the system model of Secion 2. Let $\nu_t^{(single)}$ be the probability measure of the server state at time t defined on $\mathcal{U}$. For $\phi \in \mathcal{C}_b^1(\mathcal{U})$, it can be verified that the

Kolmogorov equations are given by,

$$\begin{aligned}\langle \nu_t^{(single)}, \phi\rangle = \langle \nu_0^{(single)}, \phi\rangle + \int_{s=0}^{t} \langle \nu_s^{(single)}, \nabla_1 \phi\rangle \, ds \\ - \int_{s=0}^{t} \Bigg(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right) \, d\nu_s^{(single)}(\mathbf{x}_n) \\ + \Bigg[\left(\alpha_0 \nu_s^{(single)}(\{0\}) \left(\phi(1,0) - \phi(0)\right)\right) + \sum_{n=1}^{C-1} \sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \\ \times \alpha_n (\phi(\mathbf{x}_n^j; 0) - \phi(\mathbf{x}_n)) \, d\nu_s^{(single)}(\mathbf{x}_n) \Bigg] \Bigg) ds. \quad (14)\end{aligned}$$

On comparing the mean-field equation (9) with the Kolomogorov equation of a single-server system given by (14), it is clear that both the dynamics are similar except that α_n in equation (14) is replaced by $\lambda \Phi_n(\overline{\eta}_s)$ when the probability measure of the server state is $\overline{\eta}_s$ at time s. Equation (9) only differs from the equation with α_n in that the arrival rates depend on $\overline{\eta}_t$. This is an example of a non-linear Markov process which means that the generator of the Markov process at time t depends on the current distribution $\overline{\eta}_t$ of the Markov process [27] while in equation (14) for fixed $(\alpha_i, 0 \leq i \leq C)$ denotes a Markov process whose generator does not depend on the current distribution.

We now study the fixed-point of the mean-field. Let $P_t(0)$ be equal to $\nu_t(\{0\})$ and let $p_t(\mathbf{x}_n)$ be the probability density of ν_t $w.r.t.$ Lebesgue measure at $\mathbf{x}_n$. We obtain the differential equations satisfied by the process $(P_t, t \geq 0)$ with $P_t = (P_t(\mathbf{u}), \mathbf{u} \in \mathcal{U})$ where

$$P_t(\mathbf{y}_n) = \int_{x_1=0}^{y_1} \cdots \int_{x_n=0}^{y_n} p_t(\mathbf{x}_n) \, dx_1 \cdots dx_n. \quad (15)$$

Here, from Remark 5.2, since $\overline{\eta}_t$ is the distribution of a server's state as $N \to \infty$, it implies that $P_t(\mathbf{y}_n)$ is the probability that a server has n jobs and the i^{th} job's age is at most y_i for $1 \leq i \leq n$ as $N \to \infty$. Also, since $\overline{\eta}_t^N(\cdot) \Rightarrow \overline{\eta}(\cdot)$, for a large value of N, the fraction of servers with n jobs and the i^{th} job's age is at most y_i for $1 \leq i \leq n$ can be approximated by $P_t(\mathbf{y}_n)$.

Lemma 5.3. *The process* $(P_t, t \geq 0)$ *satisfies*

$$\frac{dP_t(0)}{dt} = \int_{y=0}^{\infty} \beta(y) \left(\frac{\partial P_t(1,y)}{\partial y}\right) dy - \lambda \Phi_0(P_t) P_t(0), \quad (16)$$

for $1 \leq n \leq C-1$,

$$\frac{dP_t(\mathbf{y}_n)}{dt} = -\sum_{i=1}^{n} \frac{\partial P_t(\mathbf{y}_n)}{\partial y_i} + \sum_{j=1}^{n+1} \int_{x_j=0}^{\infty} \beta(x_j) \left(\frac{\partial P_t(\mathbf{y}_n^j; x_j)}{\partial x_j} \right) dx_j$$
$$- \sum_{j=1}^{n} \int_{x_j=0}^{y_j} \beta(x_j) \left(\frac{\partial P_t(\mathbf{y}_n^{-j}; x_j)}{\partial x_j} \right) dx_j$$
$$+ \sum_{j=1}^{n} \lambda \frac{\Phi_{n-1}(P_t)}{n} P_t(\mathbf{y}_n^{-j}) - \lambda \Phi_n(P_t) P_t(\mathbf{y}_n), \quad (17)$$

and for $n = C$,

$$\frac{dP_t(\mathbf{y}_n)}{dt} = -\sum_{i=1}^{n} \frac{\partial P_t(\mathbf{y}_n)}{\partial y_i} - \sum_{j=1}^{n} \int_{x_j=0}^{y_j} \beta(x_j) \left(\frac{\partial P_t(\mathbf{y}_n^{-j}; x_j)}{\partial x_j} \right) dx_j$$
$$+ \sum_{j=1}^{n} \lambda \frac{\Phi_{n-1}(P_t)}{n} P_t(\mathbf{y}_n^{-j}), \quad (18)$$

where $\Phi_n(P_t) = \frac{(R_n(P_t)^d - R_{n+1}^d(P_t))}{(R_n(P_t) - R_{n+1}(P_t))}$ *and* $R_n(P_t) = \sum_{j:j=n}^{C} \lim_{b\to\infty} P_t(j, b, \cdots, b)$.

The proof is given in Appendix E.

Remark 5.3. Specializing the results to the exponential case with mean $\frac{1}{\mu}$, $\beta(x) = \mu$, and denoting $Q_t(n) = \lim_{b\to\infty} P_t(n, b, \cdots, b)$, it can be verified that the process $(Q_t, t \geq 0) = (Q_t(n), 0 \leq n \leq C, t \geq 0)$ is the unique solution of the mean-field equations given in [35] for the case of the exponential distributions with rate $\mu = 1$.

We next state the the principal result on the insensitivity of the fixed point of the MFEs. The proof is given in Section 7.

Theorem 5.3. *The process* $(P_t, t \geq 0) = (P_t(\mathbf{u}), \mathbf{u} \in \mathcal{U}, t \geq 0)$ *has a unique fixed-point given by* $\boldsymbol{\pi} = (\pi(\mathbf{y}), \mathbf{y} \in \mathcal{U})$ *where*

$$\pi(\mathbf{y}_n) = \pi_n^{(exp)} \mu^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i \quad (19)$$

and $\boldsymbol{\pi}^{(exp)} = (\pi_n^{(exp)}, 0 \leq n \leq C)$ *denotes the unique fixed-point of the mean-field when the service times are exponentially distributed with the mean* $\frac{1}{\mu}$ *and* $\pi_n^{(exp)}$ *is the stationary probability that there are* n *jobs in the limiting system. Further, since* $\int_{x=0}^{\infty} \overline{G}(x)\, dx = \frac{1}{\mu}$, *the fixed-point of the mean-field is insensitive, i.e.,*

$$\lim_{b\to\infty} \pi(n, b, \cdots, b) = \pi_n^{(exp)}. \quad (20)$$

6. Convergence of the normalized processes: proof of Theorem 5.2

By using the results on the construction a martingale in Appendix D, we now show that the normalized process $(\overline{\eta}_t^N, t \geq 0)$ converges to the mean-field limit.

Let $(\overline{\mathcal{F}}_t^N, t \geq 0)$ be the right continuous filtration associated with the process $(\overline{\eta}_t^N, t \geq 0)$. Note that we have $(\overline{\eta}_t^N, t \geq 0) \in \mathcal{D}_{\mathcal{M}_1^N(\mathcal{U})}([0,\infty))$. We first show that the sequence of processes $(\overline{\eta}_t^N, t \geq 0)$ is relatively compact and we then prove that every limit point $(\boldsymbol{\chi}_t, t \geq 0)$ almost surely has continuous sample paths with respect to t and coincide with the unique mean-field solution with the initial point ϑ. For every limit point $(\boldsymbol{\chi}_t, t \geq 0)$, $\boldsymbol{\chi}_0$ almost surely coincides with the measure ϑ from the Assumption 5.1. Further, we have that the mean-field solution is unique for the given initial measure. Hence, we conclude that for all the limit points, almost surely sample paths coincide with the unique mean-field solution $(\overline{\eta}_t, t \geq 0)$ with the initial point ϑ. The process $(\overline{\eta}_t, t \geq 0)$ is referred to as the mean-field limit. Therefore $(\overline{\eta}_t^N, t \geq 0)$ converges in distribution to the mean-field limit.

For $\phi \in \mathcal{C}_b^1(\mathcal{U})$, from Proposition D.1, the process $(\overline{M}_t^N(\phi), t \geq 0)$ defined as follows is a RCLL square integrable $\overline{\mathcal{F}}_t^N-$martingale

$$\begin{aligned}\overline{M}_t^N(\phi) = \langle \overline{\eta}_t^N, \phi\rangle - \langle \overline{\eta}_0^N, \phi\rangle - \int_{s=0}^t \langle \overline{\eta}_s^N, \nabla_1 \phi\rangle\, ds - \int_{s=0}^t \Bigg(\sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j) \\ \times \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right)\, d\overline{\eta}_s^N(\mathbf{x}_n) \\ + \Bigg[\overline{\eta}_s^N(\{0\})\lambda\Phi_0(\overline{\eta}_s^N)\left(\phi(1,0)-\phi(0)\right) + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1}\int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)} \\ \times \lambda\Phi_n(\overline{\eta}_s^N)(\phi(\mathbf{x}_n^j;0)-\phi(\mathbf{x}_n))\, d\overline{\eta}_s^N(\mathbf{x}_n)\Bigg]\Bigg) ds. \quad (21)\end{aligned}$$

We further have

$$\begin{aligned}< \overline{M}_\cdot^N(\phi) >_t = \frac{1}{N}\Bigg[\int_{s=0}^t \Bigg(\sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j)\left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right)^2\, d\overline{\eta}_s^N(\mathbf{x}_n) \\ + \Bigg[\overline{\eta}_s^N(\{0\})\lambda\Phi_0(\overline{\eta}_s^N)\left(\phi(1,0)-\phi(0)\right)^2 \\ + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1}\int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)}\lambda\Phi_n(\overline{\eta}_s^N)(\phi(\mathbf{x}_n^j;0)-\phi(\mathbf{x}_n))^2\, d\overline{\eta}_s^N(\mathbf{x}_n)\Bigg]\Bigg) ds\Bigg]. \quad (22)\end{aligned}$$

Since the space $\mathcal{D}_{\mathcal{M}_1(\mathcal{U})}([0,\infty))$ endowed with the Skorohod topology is complete and separable, by using the Prohorov's theorem [4], establishing the relative compactness of the

sequence of the processes $\{(\overline{\eta}_t^N, t \geq 0)\}$ is equivalent to proving the tightness of the processes $\{(\overline{\eta}_t^N, t \geq 0)\}$. From Theorem 4.6 of [21], Jakubowski's criteria that we recall below can be used to establish the relative compactness of the sequence of the processes $\{(\overline{\eta}_t^N, t \geq 0)\}$.

Jakubowski's criteria: A sequence of $\{X^N\}$ of $\mathcal{D}_{\mathcal{M}_1(\mathcal{U})}([0,\infty))-$ valued random elements defined on $(\Omega, \mathbb{F}, \mathbb{P})$ is tight if and only if the following two conditions are satisfied:

J1: For each $T > 0$ and $\gamma > 0$, there exists a compact set $\mathbb{K}_{T,\gamma} \subset \mathcal{M}_1(\mathcal{U})$ such that

$$\liminf_{N\to\infty} \mathbb{P}(X_t^N \in \mathbb{K}_{T,\gamma} \forall t \in [0,T]) > 1 - \gamma. \tag{23}$$

This condition is called the compact-containment condition.

J2: There exists a family $\mathcal{Q}$ of real valued continuous functions F defined on $\mathcal{M}_1(\mathcal{U})$ that separates points in $\mathcal{M}_1(\mathcal{U})$ and is closed under addition such that for every $F \in \mathcal{Q}$, the sequence $\{(F(X_t^N), t \geq 0)\}$ is tight in $\mathcal{D}_{\mathbb{R}}([0,\infty))$.

To prove the condition J2, we define a class of functions $\mathcal{Q}$ as follows:

$$\mathcal{Q} \triangleq \{F : \exists f \in \mathcal{C}_b^1(\mathcal{U}) \text{ such that } F(\nu) = \langle \nu, f \rangle, \forall \nu \in \mathcal{M}_1(\mathcal{U})\}. \tag{24}$$

Clearly every function $F \in \mathcal{Q}$ is continuous w.r.t. the weak topology on $\mathcal{M}_1(\mathcal{U})$ and further the class of functions $\mathcal{Q}$ separates points in $\mathcal{M}_1(\mathcal{U})$ and also closed under addition. We next recall the following result (From Theorem $C.9$, [38]) to prove the condition J2.

Tightness in $\mathcal{D}_{\mathbb{R}}([0,T])$: If $S = \mathcal{D}_{\mathbb{R}}([0,T])$ and $\{\mathbb{P}_n\}$ is a sequence of probability distributions on S, then $\{\mathbb{P}_n\}$ is tight if for any $\epsilon > 0$,

C1: There exists $b > 0$ such that

$$\mathbb{P}_n(|X(0)| > b) \leq \epsilon \tag{25}$$

for all $n \in \mathbb{Z}_+$.

C2: For any $\gamma > 0$, there exists $\rho > 0$ such that

$$\mathbb{P}_n(w_X(\rho) > \gamma) \leq \epsilon \tag{26}$$

for n sufficiently large, where

$$w_X(\rho) = \sup\{|X(t) - X(s)| : s, t \leq T, |s - t| \leq \rho\} \tag{27}$$

and any limiting point $\mathbb{P}$ satisfies $\mathbb{P}(\mathcal{C}_{\mathbb{R}}([0,T])) = 1$.

We first establish the relative compactness of the sequence $\{(\overline{\eta}_t^N, t \geq 0)\}$. For this, we next prove the conditions C1 and C2 that are sufficient to prove the relative compactness of the sequence $\{(\langle\overline{\eta}_t^N, \phi\rangle, t \geq 0)\}$ for $\phi \in \mathcal{C}_b^1(\mathcal{U})$ in $D_{\mathbb{R}}([0,\infty))$. For any $T > 0$, $t \in [0,T]$, we have $\langle\overline{\eta}_t^N, \phi\rangle \leq \|\phi\|_1 \langle\overline{\eta}_t^N, \mathbf{1}\rangle$ and since $\langle\overline{\eta}_t^N, \mathbf{1}\rangle = 1$, the condition C1 is trivially satisfied with $b = \|\phi\|_1$.

We next prove that the condition C2 holds. For $\epsilon > 0$, by using equation (22) and the Doob's inequality [13, page 63], we have

$$\begin{aligned}\mathbb{P}\left(\sup_{t\leq T}\left|\overline{M}_t^N(\phi)\right| \geq \epsilon\right) &\leq \frac{4}{\epsilon^2}\mathbb{E}\left[< \overline{M}_\cdot^N(\phi) >_T\right] \\ &\leq 4T\|\phi\|^2\frac{1}{N}(\|\beta\| + d\lambda)\end{aligned}$$

and hence, $\mathbb{P}\left(\sup_{t\leq T}\left|\overline{M}_t^N(\phi)\right| \geq \epsilon\right) \to 0$ as $N \to \infty$. Therefore the sequence of processes $\{(\overline{M}_t^N(\phi), t \geq 0)\}$ converges in distribution to the null process from the standard convergence criterion in $\mathcal{D}_{\mathbb{R}}([0,T])$. Further, the sequence of processes $\{(\overline{M}_t^N(\phi), t \geq 0)\}$ is tight in $\mathcal{D}_{\mathbb{R}}([0,T])$ and hence, there exists $\rho' > 0$ and $N' > 0$ such that for all $N \geq N'$, we have

$$\mathbb{P}\left(\sup_{u,v\leq T, |u-v|\leq\rho'}\left|\overline{M}_v^N(\phi) - \overline{M}_u^N(\phi)\right| \geq \frac{\gamma}{2}\right) \leq \frac{\epsilon}{2} \tag{28}$$

For any $u < v \leq T$, from equation (21), we have

$$\begin{aligned}\left|\langle\overline{\eta}_v^N, \phi\rangle - \langle\overline{\eta}_u^N, \phi\rangle\right| \leq \int_{s=u}^{v}\left|\langle\overline{\eta}_s^N, \nabla_1\phi\rangle\right| ds + 2\|\beta\|\|\phi\|C\,|u-v| + 2\|\phi\|\lambda\,|u-v| \\ + \left|\overline{M}_v^N(\phi) - \overline{M}_u^N(\phi)\right|. \end{aligned} \tag{29}$$

Further, we can write

$$\left|\langle\overline{\eta}_v^N, \phi\rangle - \langle\overline{\eta}_u^N, \phi\rangle\right| \leq |v-u|\,C\|\phi\|_1(1 + 2\|\beta\| + 2d\lambda) + \left|\overline{M}_v^N(\phi) - \overline{M}_u^N(\phi)\right|. \tag{30}$$

Therefore by using equations (28) and (30), there exists $\rho > 0$ and $N_1 > 0$ such that for $N \geq N_1$, we have $\mathbb{P}\left(\sup_{u,v\leq T, |u-v|\leq\rho}\left|\langle\overline{\eta}_v^N, \phi\rangle - \langle\overline{\eta}_u^N, \phi\rangle\right| \geq \gamma\right) \leq \epsilon$. This proves the condition C2. Since the conditions C1 and C2 hold, the condition J2 also holds.

We next prove the compact containment condition J1. Let $(n_i(t), x_{i1}(t)\ldots, x_{in_i(t)}(t))$ be the state of the i^{th} server at time t where $x_{ij}(t)$ denotes the age of the j^{th} job at the i^{th} server. Clearly, we have $\langle\overline{\eta}_t^N, \mathcal{I}\rangle = \frac{1}{N}\sum_{i=1, n_i(t)>0}^{N}(x_{i1}(t) + \cdots + x_{in_i(t)}(t))$.

We can classify the progressing jobs into two classes. The jobs that are in service from the beginning ($t = 0$) form the first class and the second class of jobs are the ones that entered

the system in the interval $(0, t]$. At a server, the number of progressing jobs that belong to each class are upper bounded by C. Let Y_t be a random variable representing the age of a job belonging to the second class that is in progress at time t, and Y be a random variable with job length distribution G, then for any $b \geq 0$, we have

$$\mathbb{P}(Y_t \geq b) \leq \mathbb{P}(Y \geq b). \tag{31}$$

Therefore, using equation (31), since each server has capacity C, for any time $t \geq 0$, we can write

$$\mathbb{P}(\langle \overline{\eta}_t^N, \mathcal{I} \rangle \geq b) \leq \mathbb{P}\left(\langle \overline{\eta}_0^N, \mathcal{I} \rangle + Ct + \frac{1}{N} \sum_{i=1}^{N} (Y_{i1} + \ldots + Y_{iC}) \geq b \right), \tag{32}$$

where $(Y_{ij}, 1 \leq i \leq N, 1 \leq j \leq C)$ are i.i.d random variables with distribution G. Further, by weak law of large numbers, we have $\frac{1}{N} \sum_{i=1}^{N} (Y_{i1} + \ldots + Y_{iC}) \Rightarrow \frac{C}{\mu}$ as $N \to \infty$. Therefore, by choosing $Z_T = 2\langle \vartheta, \mathcal{I} \rangle + 2CT + \frac{2C}{\mu}$, we have

$$\mathbb{P}\left(\sup_{t \in [0,T]} \langle \overline{\eta}_t^N, \mathcal{I} \rangle > Z_T \right) \to 0 \tag{33}$$

as $N \to \infty$. Let us define $\mathcal{L}_T \triangleq \{\zeta \in \mathcal{M}_1(\mathcal{U}) : \langle \zeta, \mathcal{I} \rangle \leq Z_T\}$. Since $\langle \zeta, \mathcal{I} \rangle \leq Z_T$ for $\zeta \in \mathcal{L}_T$, let $B = \mathcal{U}_0 \cup (\cup_{n \geq 1} \{(n, y_1, \cdots, y_n) : 0 \leq y_i \leq r, 1 \leq i \leq n\})$ and $\overline{B}$ be the compliment of B, then we have $\zeta(\overline{B}) \leq \frac{Z_T}{r}$. Hence, $\lim_{r \to \infty} \sup_{\zeta \in \mathcal{L}_T} \zeta(\overline{B}) = 0$. Therefore from Lemma A7.5 of [23], $\mathcal{L}_T$ is relatively compact in $\mathcal{M}_1(\mathcal{U})$. Further, from equation (33), we have $\liminf_{N \to \infty} \mathbb{P}(\overline{\eta}_t^N \in \mathcal{L}_T, \forall t \in [0, T]) > 1 - \gamma$. Let $\mathbb{K}_T$ be the closure of $\mathcal{L}_T$, then we have a compact set $\mathbb{K}_T$ such that $\liminf_{N \to \infty} \mathbb{P}(\overline{\eta}_t^N \in \mathbb{K}_T, \forall t \in [0, T]) > 1 - \gamma$ for all $0 < \gamma < 1$.

This establishes the condition J1 and hence the proof of the tightness of the sequence of processes $(\overline{\eta}_t^N, t \geq 0)$ is completed.

Let $(\boldsymbol{\chi}_t, t \geq 0)$ be a limit of a converging subsequence $\{(\overline{\eta}_t^{N_{i_k}}, t \geq 0)\}$. From the condition C2, χ_t is continuous in t , $\mathbf{P}_\chi - a.s.$, where $\mathbf{P}_\chi$ is the probability law of $(\chi_t, t \geq 0)$. Furthermore from [22, Theorem 1.7] for $f \in \mathcal{C}_b(\mathcal{U}), \nu \in \mathcal{M}_1(\mathcal{U})$, it follows that for any $T > 0$, we have $(\nu_t, 0 \leq t \leq T) \mapsto (\langle \nu_t, f \rangle, 0 \leq t \leq T)$ is continuous in the Skorohod topology. Then since the martingale $(\overline{M}_t^{N_{i_k}}(\phi), t \geq 0)$ converges to the null process, by the

continuous mapping theorem, we conclude

$$\langle \chi_t, \phi\rangle = \langle \chi_0, \phi\rangle + \int_{s=0}^{t} \langle \chi_s, \nabla_1 \phi\rangle \, ds \\ - \int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n) \right) \, d\chi_s(\mathbf{x}_n) \right. \\ + \left[(\chi_s(\{0\}) \lambda \Phi_0(\chi_s) \left(\phi(1,0) - \phi(0) \right)) + \sum_{n=1}^{C-1} \sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \right. \\ \left. \left. \times \lambda \Phi_n(\chi_s) (\phi(\mathbf{x}_n^j; 0) - \phi(\mathbf{x}_n)) \, d\chi_s(\mathbf{x}_n) \right] \right) ds. \quad (34)$$

From the Assumption 5.1, $\boldsymbol{\chi}_0 = \vartheta$ almost surely and hence the sample paths coincide almost surely with the unique mean-field solution with the initial point ϑ. This argument holds for every limit point, and hence, the sample paths of every limit point are almost surely the same as the deterministic mean-field solution with the initial point ϑ. This completes the proof.

7. Insensitivity: proof of Theorem 5.3

We now show that $\boldsymbol{\pi} = (\pi(\mathbf{u}), \mathbf{u} \in \mathcal{U})$ is the unique fixed-point of the mean-field. From [35], we first recall that under the assumption of exponential service time distributions, there exists a unique probability measure of occupancy $\boldsymbol{\pi}^{(exp)} = (\pi_n^{(exp)}, 0 \leq n \leq C)$ on $\{0, 1, \cdots, C\}$ to the stationary MFEs given below,

$$\lambda_n^{(exp)}(\boldsymbol{\pi}^{(exp)}) \pi_n^{(exp)} = (n+1) \mu \pi_{n+1}^{(exp)}, \quad (35)$$

where

$$\lambda_n^{(exp)}(\boldsymbol{\pi}^{(exp)}) = \lambda \frac{(\sum_{j=n}^{C} \pi_j^{(exp)})^d - (\sum_{j=n+1}^{C} \pi_j^{(exp)})^d}{(\sum_{j=n}^{C} \pi_j^{(exp)}) - (\sum_{j=n+1}^{C} \pi_j^{(exp)})}. \quad (36)$$

Let $\theta = (\theta(\mathbf{u}), \mathbf{u} \in \mathcal{U})$ be a fixed-point of the MFEs of the process $(P_t, t \geq 0)$ under general service time distributions. Using $\boldsymbol{\theta}$, let the corresponding probability measure of occupancy be $\boldsymbol{\Gamma} = (\Gamma_n, 0 \leq n \leq C)$ defined such that $\Gamma_n = \lim_{b \to \infty} \theta(n, b, \cdots, b))$ and $\Gamma_0 = \theta(0)$. We now show that

$$\theta(\mathbf{y}_n) = \frac{\left(\prod_{i=1}^{n} \frac{\lambda_{i-1}^{(exp)}(\boldsymbol{\Gamma})}{i\mu} \right)}{1 + \sum_{m=1}^{C} \left(\prod_{i=1}^{m} \frac{\lambda_{i-1}^{(exp)}(\boldsymbol{\Gamma})}{i\mu} \right)} \mu^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i) \, dx_i \quad (37)$$

and

$$\theta(0) = \frac{1}{1 + \sum_{m=1}^{C} \left(\prod_{i=1}^{m} \frac{\lambda_{i-1}^{(exp)}(\mathbf{\Gamma})}{i\mu} \right)}. \tag{38}$$

Then it implies that $\mathbf{\Gamma}$ also satisfies equations (35)-(36), and hence $\mathbf{\Gamma} = \boldsymbol{\pi}^{(exp)}$ concluding the insensitivity of the fixed-point. Furthermore, we have that $\theta(\mathbf{y}_n) = \pi_n^{(exp)} \mu^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i$ concluding the uniqueness of the fixed-point of the mean-field under general service time distributions.

To complete the proof, it remains to show the validity of equations (37)-(38). We now recall the stationary distribution $\boldsymbol{\pi}^{(single)} = (\pi^{(single)}(\mathbf{u}), \mathbf{u} \in \mathcal{U})$ of a single server loss system with state-dependent Poisson arrival process with rate α_n $(0 \leq n \leq C)$ when there are n progressing jobs and the service time distributions are as in the system model of Section 2. Then from [9], the stationary probability that the server has n progressing jobs and the i^{th} job has age at most y_i $(1 \leq i \leq n)$ is given by

$$\pi^{(single)}(\mathbf{y}_n) = \frac{\left(\prod_{i=1}^{n} \frac{\alpha_{i-1}}{i\mu} \right)}{1 + \sum_{m=1}^{C} \left(\prod_{i=1}^{m} \frac{\alpha_{i-1}}{i\mu} \right)} \mu^n \prod_{i=1}^{n} \int_{x_i=0}^{y_i} \overline{G}(x_i)\, dx_i \tag{39}$$

and

$$\pi^{(single)}(0) = \frac{1}{1 + \sum_{m=1}^{C} \left(\prod_{i=1}^{m} \frac{\alpha_{i-1}}{i\mu} \right)}. \tag{40}$$

For the given fixed-point $\boldsymbol{\theta}$ of the mean-field and its corresponding occupancy probability measure $\mathbf{\Gamma}$, consider a single server system under the assumption of a Poisson arrival process with state-dependent rate $\lambda_n^{(exp)}(\mathbf{\Gamma})$ $(0 \leq n \leq C)$ when there are n progressing jobs. Then the unique stationary distribution is given by equations (39)-(40) with α_n replaced by $\lambda_n^{(exp)}(\mathbf{\Gamma})$ for all $0 \leq n \leq C$. But from equations (9), (14), and Lemma 5.3, since $\overline{R}_n(\boldsymbol{\theta}) = \sum_{j=n}^{C} \Gamma_j$, we have that $\boldsymbol{\theta}$ is also an another stationary distribution for the single server system with state dependent Poisson arrival process having rates $\lambda_n^{(exp)}(\mathbf{\Gamma})$ for all $0 \leq n \leq C$. Since the stationary distribution must be unique, equations (37)-(38) must hold. This completes the proof.

8. Numerical results

Showing that the fixed-point of the mean-field approximates the stationary distribution of the system with large N, remains an open problem. If one can establish that the equilibrium

or fixed-point of the MFEs is globally asymptotically stable (GAS), then the conclusion of the interchange of limits would follow from the Prohorov's theorem [4]. Proving that the equilibrium point of the MFEs is GAS is a challenging problem because the joint distribution of the occupancy and ages does not possess any monotonicity properties unlike the case of exponential service time distributions [35]. In this section, we present numerical results on the validity of the GAS of the mean-field for the case in which the service time distributions are mixed-Erlang. In this case, the state of a server is also multi-dimensional and the mean-field is also not monotonic unlike the exponential case. It is numerically easier to solve the MFEs for the case of mixed-Erlang distributions as they are systems of ODEs unlike the case of general service time distributions for which the MFEs are PDEs as we have shown. One more reason for using mixed-Erlang distributions is that such distributions are dense in the set of all distributions that have support on $\mathbb{R}_+$, see [3]. Our numerical results show that the mean-field is GAS for the case of mixed-Erlang service time distributions.

We consider the system parameters as follows: The capacity of a server is assumed to be $C = 5$. The average job length is assumed to be equal to one, i.e. $\mu = 1$. The service times have a Mixed-Erlang distribution given by sums of independent exponentially distributed random variables (known as an Erlang distribution) where the number of exponential phases (or independent random exponentials) is equal to $i \in \{1, 2, \ldots, M\}$ with probability p_i such that $\sum_{i=1}^{M} p_i = 1$. Each exponential phase is assumed to have rate μ_p. Therefore, we have,

$$\frac{1}{\mu} = \frac{\sum_{i=1}^{M} i p_i}{\mu_p}.$$

We choose $M = 3, p_1 = .3, p_2 = 0.3, p_3 = 0.4$.

Under mixed-Erlang service time distribution assumptions, let S be the set of all possible server states defined as $S = \cup_{n=0}^{C} S_n$ where $S_0 = \{(0)\}$ and $S_n = \{(n, r_1, \ldots, r_n) : 1 \leq r_i \leq M, 1 \leq i \leq n\}$. We refer to an element in the set S by $\mathbf{r}$ and an element in the set S_n by $\mathbf{r}_n$. The system dynamics can be modeled as a Markov process $\mathbf{x}^N(t) = (x_{\mathbf{r}}^N(t), \mathbf{r} \in S)$ where $x_{\mathbf{r}}^N(t)$ denotes the fraction of servers with n jobs such that i^{th} job has r_i remaining phases at time t. Since the Markov process $(\mathbf{x}^N(t), t \geq 0)$ is defined on a finite dimensional space, we can establish the mean-field limit $\mathbf{x}(t) = (x_{\mathbf{r}}(t), \mathbf{r} \in S)$ by using the same procedure as that of the exponential service times case in [35]. Hence we recall the following result without proof from [44].

Proposition 8.1. *If $\mathbf{x}^N(0)$ converges in distribution to a state $\mathbf{u}$, then the process $\mathbf{x}^N(\cdot)$ converges in distribution to a deterministic process $\mathbf{x}(\cdot,\mathbf{u})$ as $N\to\infty$ called the mean-field. The process $\mathbf{x}(\cdot,\mathbf{u})$ is the unique solution of the following system of differential equations.*

$$\mathbf{x}(0,\mathbf{u}) = \mathbf{u}, \tag{41}$$

$$\dot{x}_{\mathbf{r}_n}(t,\mathbf{u}) = h_{\mathbf{r}_n}(\mathbf{x}(t,\mathbf{u})), \tag{42}$$

and $\mathbf{h} = (h_{\mathbf{r}}, \mathbf{r}\in S)$ with the mapping $h_{\mathbf{r}_n}$ given by

$$\begin{aligned} h_{\mathbf{r}_n}(\mathbf{x}) = &\sum_{b=1}^{n}\left(\frac{p_{r_b}}{n}\right) x_{(\mathbf{r}_n^{-j})}\lambda_{n-1}^{(ME)}(\mathbf{x}) - x_{\mathbf{r}_n}\lambda_n^{(ME)}(\mathbf{x})I_{\{n<C\}} \\ &+\sum_{b=1}^{n+1}\mu_p I_{\{n<C\}}x_{(\mathbf{r}_n^b;1)} + \sum_{b=1}^{n}\mu_p x_{(n,r_1,\cdots,r_{b-1},r_b+1,r_{b+1},\cdots,r_n)} - n\mu_p x_{\mathbf{r}_n}, \end{aligned} \tag{43}$$

where

$$\lambda_n^{(ME)}(\mathbf{u}) = \frac{\lambda}{\left(\sum_{\mathbf{r}_n} u_{\mathbf{r}_n}\right)}\left[\left(\sum_{i=n}^{C}\sum_{\mathbf{b}_i} u_{\mathbf{b}_i}\right)^d - \left(\sum_{i=n+1}^{C}\sum_{\mathbf{b}_i} u_{\mathbf{b}_i}\right)^d\right]. \tag{44}$$

In Figure 1, we plot $d_E^2(\mathbf{x}(t,\mathbf{u}),\pi)$ as a function of t where d_E is the euclidean distance defined by

$$d_E(\mathbf{u},\mathbf{v}) = \sqrt{\sum_{\mathbf{l}\in S}|u_{\mathbf{l}} - v_{\mathbf{l}}|^2}.$$

It is observed that for $d=2$, $\lambda=1$, and for four different initial points $\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3$, and $\mathbf{u}_4$, the mean-field $\mathbf{x}(t,\mathbf{u})$ for mixed-Erlang service time distribution converges to its unique fixed-point $\boldsymbol{\pi}$. Note that the computed $\boldsymbol{\pi}$ depends on the chosen value of d. This supports that $\boldsymbol{\pi}$ is globally stable.

We conclude with some numerical results for the blocking probability of the above system showing closeness to the theoretical lower bound. Under asymptotic independence any finite set of servers are independent and the fixed-point of the mean-field implies that the fixed point is the stationary distribution of the state of a server. The average blocking probability is then given by $\pi_q(C)^d$ where $\pi_q(C) = \lim_{b\to\infty}\pi(C,b,\cdots,b)$. Let us recall the lower bound on the average blocking probability, denoted by P_{block}^{avg} for any routing scheme shown in [36]. From the Little's law, the average number of customers in the system is equal to $(1-P_{block}^{avg})N\lambda$ which is upper bounded by NC. Hence,

$$P_{block}^{avg} \geq \left(1-\frac{C}{\lambda}\right)_+,$$

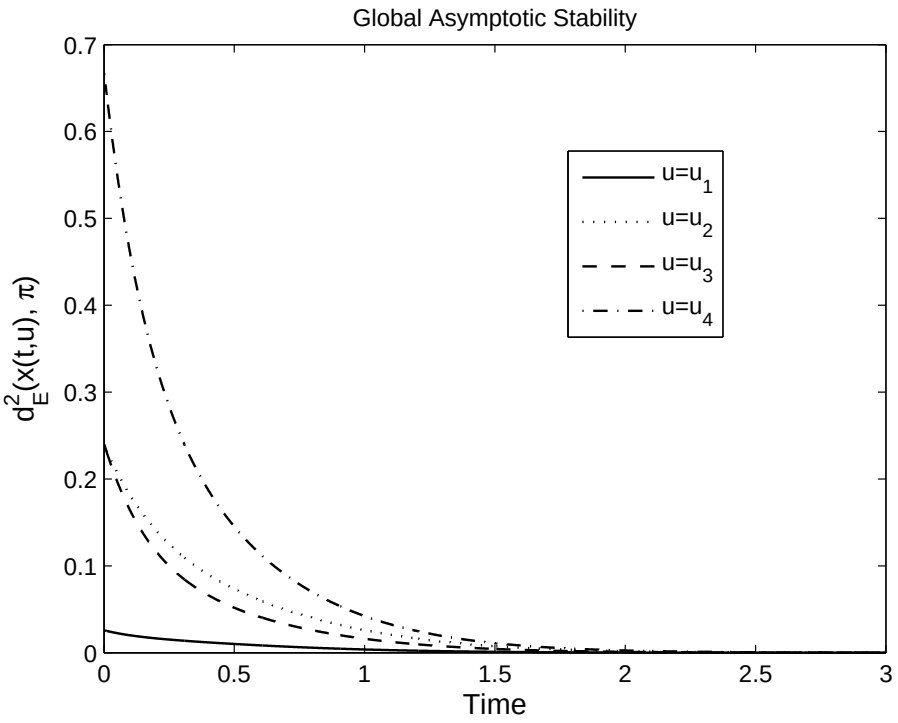

FIGURE 1: Convergence of mean-field to the fixed-point

where $(x)_+ = \max(x, 0)$. In Figure 2, we plot the lower bound $\left(1 - \frac{C}{\lambda}\right)_+$ and the average blocking probability under the SQ(d) routing, and the state-independent random routing where a destination server is chosen uniformly at random, as a function of λ. It is clear that the resulting average blocking probability under the SQ(d) policy is much lower than the resulting average blocking probability when pure random routing is employed. Furthermore, the average blocking probability under the SQ(d) routing approaches the lower bound as d increases.

9. Concluding Remarks

.

In this paper we have provided a measure-valued process approach to establish the mean-field behavior of loss systems with SQ(d) routing and general service time requirements. The extension of these results to multi-class systems where servers are classified into different classes based on their capacities and jobs are classified into different classes based on their

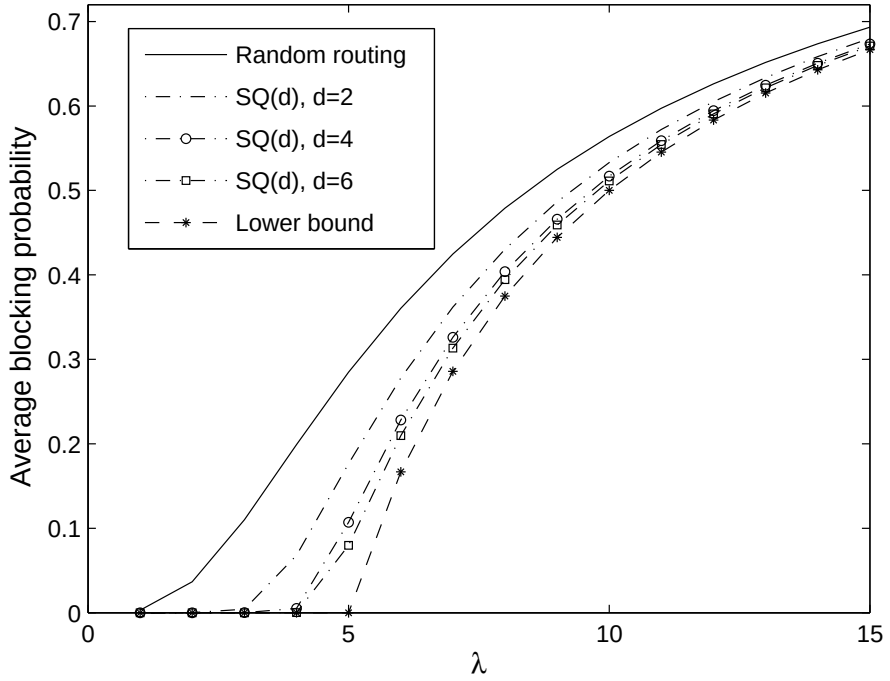

FIGURE 2: Comparison of the average blocking probability under SQ(d) with lower bound.

service requirements follow in a similar manner *mutatis mutandis* from the approach used here. Establishing the global asymptotic stability of the unique fixed point remains an open problem.

Appendix A. Properties of Markov process and its semigroup

In this section, we compute the semigroup of the Markov process $(\eta_t^N, t \geq 0)$ and we then show that the Markov process $(\eta_t^N, t \geq 0)$ is a Feller process.

Let A_h be the number of arrivals in the interval $[0, h]$. Similarly, given the initial state η_0^N, let D_h be the number of departures that occur in the interval $[0, h]$. Note that a job with age x at time t departs from the system in the interval $[t, t + h]$ with the probability $\frac{G(x+h)-G(x)}{\overline{G}(x)}$. Further, from the definition of the hazard rate, we have that $\lim_{h\to 0} \frac{1}{h} \frac{G(x+h)-G(x)}{\overline{G}(x)} = \beta(x)$ and hence

$$\frac{G(x+h) - G(x)}{\overline{G}(x)} = \beta(x)h + o(h). \tag{45}$$

Let $\mathcal{F}_t^N$ be the filtration:

$$\mathcal{F}_t^N = \cap_{\epsilon>0}\sigma(\{\eta_s^N : s \leq t+\epsilon\}). \tag{46}$$

We now define

$$T_h^N f(\nu) = \mathbb{E}\left[f(\eta_h^N)|\eta_0^N = \nu\right]$$

where f is a continuous bounded function $f : \mathcal{M}_F(\mathcal{U}) \to \mathbb{R}$ and the operator T_h^N is a semigroup operator when $(\eta_t^N, t \geq 0)$ is a Markov process. Before computing the expression for $T_h^N f(\nu)$, we first introduce the following notation. Suppose the measure $\eta_0^N = \nu$ has mass at m atoms and let the i^{th} atom be $v^{(i)} = (n_i, v_1^{(i)}, \cdots, v_{n_i}^{(i)})$ for $1 \leq i \leq m$ and let the number of servers with the state $v^{(i)}$ be denoted by $\nu(\{v^{(i)}\}) = a^{(i)}$. If a server lies in the state $\mathbf{b}_n = (n, b_1, \ldots, b_n)$ at time t, let the probability that there is no departure at in the interval $[t, t+h]$ be denoted by $p_{ND}(\mathbf{b}_n, h)$. We then have

$$p_{ND}(\mathbf{b}_n, h) = \prod_{i=1}^{n} \frac{\overline{G}(b_i + h)}{\overline{G}(b_i)}. \tag{47}$$

Note that using equation (45), we can write

$$p_{ND}(\mathbf{b}_n, h) = \prod_{j=1}^{n} (1 - \beta(b_j)h) + o(h). \tag{48}$$

Lemma A.1. *Let f be a real valued continuous bounded function defined on $\mathcal{M}_F(\mathcal{U})$. Then the process $(\eta_t^N, t \geq 0)$ is a weak-homogeneous $\mathcal{M}_F(\mathcal{U})$-valued Markov process with semi-*

group operator $T_h^N(\cdot)$ given by

$$T_h^N f(\nu) = (1 - N\lambda h)\left(\prod_{j=1,n_j>0}^{m}(p_{ND}(\mathbf{v}^{(j)},h))^{a^{(j)}}\right)f(\tau_h\nu)$$
$$+ (1 - N\lambda h)\sum_{j=1,n_j>0}^{m}\sum_{r=1}^{n_j}a^{(j)}\left(\frac{G(v_r^{(j)}+h)-G(v_r^{(j)})}{\overline{G}(v_r^{(j)})}\right)$$
$$\times\left(\prod_{w=1,w\neq r}^{n_j}\left(\frac{\overline{G}(v_w^{(j)}+h)}{\overline{G}(v_w^{(j)})}\right)\right)(p_{ND}(\mathbf{v}^{(j)},h))^{(a^{(j)}-1)}\left(\prod_{i=1,n_i>0,i\neq j}^{m}(p_{ND}(\mathbf{v}^{(i)},h))^{a^{(i)}}\right)$$
$$\times f(\tau_h\nu + \delta_{((\tau_h^+\mathbf{v}^{(j)})^{-r})} - \delta_{(\tau_h^+\mathbf{v}^{(i)})})$$
$$+ (N\lambda h)\int_{x=0}^{h}\frac{1}{h}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\frac{1}{n_i+1}p_r(\tau_x\nu:\mathbf{v}^{(i)})\right.$$
$$\times\left[I_{\{n_i<C\}}f\left(\tau_h\nu+\delta_{((\tau_h^+\mathbf{v}^{(i)})^j;h-x)}-\delta_{(\tau_h^+\mathbf{v}^{(i)})}\right)\prod_{k=1,n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\overline{G}(h-x)\right.$$
$$\left.\left.+I_{\{n_i=C\}}f\left(\tau_h\nu\right)\prod_{k=1,n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\right]\right)dx+\epsilon(\nu,h), \quad (49)$$

where $\epsilon(\nu,h)$ is a $o(h)$ term for all ν. Moreover, the process $(\eta_t^N, t\geq 0)$ is a Feller-Dynkin process.

Proof. We can write

$$T_h^N f(\nu) = \mathbb{E}\left[f(\eta_h^N)I_{\{A_h=0,D_h=0\}}|\eta_0^N=\nu\right]+\mathbb{E}\left[f(\eta_h^N)I_{\{A_h=0,D_h=1\}}|\eta_0^N=\nu\right]$$
$$+\mathbb{E}\left[f(\eta_h^N)I_{\{A_h=1,D_h=0\}}|\eta_0^N=\nu\right]+\sum_{i\geq 1,j\geq 1}\mathbb{E}\left[f(\eta_h^N)I_{\{A_h=i,D_h=j\}}|\eta_0^N=\nu\right]. \quad (50)$$

We first simplify the first term on the right side of equation (50). In this case, since there are no arrivals or departures, we have $\eta_h^N = \tau_h\nu$. As a consequence, we have

$$\mathbb{E}\left[f(\eta_h^N)I_{\{A_h=0,D_h=0\}}|\eta_0^N=\nu\right] = f(\tau_h\nu)\mathbb{P}\left((\{A_h=0,D_h=0\})|\eta_0^N=\nu\right). \quad (51)$$

Further, we can write

$$\mathbb{P}\left((\{A_h=0,D_h=0\})|\eta_0^N=\nu\right)=\mathbb{P}\left(\{A_h=0\}|\eta_0^N=\nu\right)\mathbb{P}\left((\{D_h=0\})|A_h=0,\eta_0^N=\nu\right).$$

Since the arrival process is a Poisson process with rate $N\lambda$ and hence, it is independent of the state ν. Therefore, we have $\mathbb{P}\left(\{A_h=0\}|\eta_0^N=\nu\right)=\mathbb{P}(\{A_h=0\})=e^{-(N\lambda h)}$. On the other

hand, the number of departures D_h, is influenced by the number of arrivals A_h. Hence we need to compute the expression of $\mathbb{P}\left((\{D_h = j\})|A_h = i, \eta_0^N = \nu\right)$ that gives the probability that there are j departures in the interval $[0, h]$ conditioned on the event that there are i arrivals in the interval $[0, h]$. As it is known, if the arrival process is a Poisson process, conditioned on the number of arrivals A_h, the arrival instants are random variables with uniform distribution in the interval $[0, h]$ [39, p. 325]. It can be seen that

$$\mathbb{P}\left((\{D_h = 0\})|A_h = 0, \eta_0^N = \nu\right) = \prod_{j=1, n_j>0}^{m} (p_{ND}(\mathbf{v}^{(j)}, h))^{a^{(j)}}.$$

We can write

$$\mathbb{E}\left[f(\eta_h^N) I_{\{A_h=0, D_h=0\}} | \eta_0^N = \nu\right] = (1 - N\lambda h)\left(\prod_{j=1, n_j>0}^{m} (p_{ND}(\mathbf{v}^{(j)}, h))^{a^{(j)}}\right) f(\tau_h \nu) + \epsilon_1(\nu, h),$$

where

$$\epsilon_1(\nu, h) = (\mathbb{P}(\{A_h = 0\}) - (1 - N\lambda h)) \prod_{j=1, n_j>0}^{m} (p_{ND}(\mathbf{v}^{(j)}, h))^{a^{(j)}} f(\tau_h \nu)$$

is a $o(h)$ term for all ν.

Similarly, we can write the second term of the right side of equation (50) as

$$\mathbb{E}\left[f(\eta_h^N) I_{\{A_h=0, D_h=1\}} | \eta_0^N = \nu\right] = (1 - N\lambda h) \sum_{j=1, n_j>0}^{m} \sum_{r=1}^{n_j} a^{(j)} \left(\frac{G(v_r^{(j)} + h) - G(v_r^{(j)})}{\overline{G}(v_r^{(j)})}\right) \times \left(\prod_{w=1, w\neq r}^{n_j} \left(\frac{\overline{G}(v_w^{(j)} + h)}{\overline{G}(v_w^{(j)})}\right)\right) (p_{ND}(\mathbf{v}^{(j)}, h))^{(a^{(j)}-1)} \left(\prod_{i=1, n_i>0, i\neq j}^{m} (p_{ND}(\mathbf{v}^{(i)}, h))^{a^{(i)}}\right) \times f(\tau_h \nu + \delta_{((\tau_h^+ \mathbf{v}^{(j)})-r)} - \delta_{(\tau_h^+ \mathbf{v}^{(i)})}) + \epsilon_2(\nu, h),$$

where we use r to denote the index of the departing job at a server with the state $\mathbf{v}^{(j)}$ and

$\epsilon_2(\nu,h)$ is a $o(h)$ term for all ν given by

$$\epsilon_2(\nu,h) = (\mathbb{P}(\{A_h=0\}) - (1-N\lambda h)) \sum_{j=1,n_j>0}^{m} \sum_{r=1}^{n_j} a^{(j)}$$
$$\times \left(\frac{G(v_r^{(j)}+h) - G(v_r^{(j)})}{\overline{G}(v_r^{(j)})}\right) \left(\prod_{w=1,w\neq r}^{n_j} \left(\frac{\overline{G}(v_w^{(j)}+h)}{\overline{G}(v_w^{(j)})}\right)\right) (p_{ND}(\mathbf{v}^{(j)},h))^{(a^{(j)}-1)}$$
$$\times \left(\prod_{i=1,n_i>0,i\neq j}^{m} (p_{ND}(\mathbf{v}^{(i)},h))^{a^{(i)}}\right) f(\tau_h\nu + \delta_{((\tau_h^+\mathbf{v}^{(j)})^{-r})} - \delta_{(\tau_h^+\mathbf{v}^{(i)})}).$$

We next compute the third term on the right side of equation (50). We can write

$$\mathbb{E}\left[f(\eta_h^N) I_{\{A_h=1,D_h=0\}} | \eta_0^N = \nu\right] = (\mathbb{P}(\{A_h=1\}))$$
$$\times \int_{x=0}^{h} \frac{1}{h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \frac{1}{n_i+1} p_r(\tau_x\nu : \tau_x^+\mathbf{v}^{(i)}) \left[I_{\{n_i<C\}} f\left(\tau_h\nu + \delta_{((\tau_h^+\mathbf{v}^{(i)})^j;h-x)} - \delta_{(\tau_h^+\mathbf{v}^{(i)})}\right)\right.\right.$$
$$\left.\left. \times \prod_{k=1,n_k>0}^{m} (p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}} \overline{G}(h-x) + I_{\{n_i=C\}} f\left(\tau_h\nu\right) \prod_{k=1,n_k>0}^{m} (p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\right]\right) dx,$$

where the arrival instant x is chosen uniformly in $[0,h]$ given $A_h=1$, i denotes the index of the atom corresponding to the state of the destination server and j is the position of the routed job at the destination server chosen uniformly at random from n_i+1 positions. Further, we write

$$\mathbb{E}\left[f(\eta_h^N) I_{\{A_h=1,D_h=0\}} | \eta_0^N = \nu\right] = (N\lambda h) \int_{x=0}^{h} \frac{1}{h} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \frac{1}{n_i+1} p_r(\tau_x\nu : \tau_x^+\mathbf{v}^{(i)})\right.$$
$$\times \left[I_{\{n_i<C\}} f\left(\tau_h\nu + \delta_{((\tau_h^+\mathbf{v}^{(i)})^j;h-x)} - \delta_{(\tau_h^+\mathbf{v}^{(i)})}\right) \prod_{k=1,n_k>0}^{m} (p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}} \overline{G}(h-x)\right.$$
$$\left.\left. + I_{\{n_i=C\}} f\left(\tau_h\nu\right) \prod_{k=1,n_k>0}^{m} (p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\right]\right) dx + \epsilon_3(\nu,h),$$

where

$$\epsilon_3(\nu,h) = (\mathbb{P}(\{A_h=1\}) - N\lambda h)\int_{x=0}^{h}\frac{1}{h}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\frac{1}{n_i+1}p_r(\tau_x\nu:\tau_x^+\mathbf{v}^{(i)})\right.$$
$$\left[I_{\{n_i<C\}}f\left(\tau_h\nu+\delta_{((\tau_h^+\mathbf{v}^{(i)})^j;h-x)}-\delta_{(\tau_h^+\mathbf{v}^{(i)})}\right)\prod_{k=1,n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\overline{G}(h-x)\right.$$
$$\left.\left.+I_{\{n_i=C\}}f\left(\tau_h\nu\right)\prod_{k=1,n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\right]\right)dx.$$

We now show that $\epsilon_3(\nu,h)$ is a $o(h)$ term for all ν. For this, we apply the method of change of variables by replacing x with hy. As a consequence, we have

$$\epsilon_3(\nu,h) = (\mathbb{P}(\{A_h=1\}) - N\lambda h)\,h\int_{y=0}^{1}\frac{1}{h}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\frac{1}{n_i+1}p_r(\tau_{hy}\nu:\tau_{hy}^+\mathbf{v}^{(i)})\right.$$
$$\left[I_{\{n_i<C\}}f\left(\tau_h\nu+\delta_{((\tau_h^+\mathbf{v}^{(i)})^j;h-hy)}-\delta_{(\tau_h^+\mathbf{v}^{(i)})}\right)\prod_{k=1,n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\overline{G}(h-hy)\right.$$
$$\left.\left.+I_{\{n_i=C\}}f\left(\tau_h\nu\right)\prod_{k=1,n_k>0}^{m}(p_{ND}(\mathbf{v}^{(k)},h))^{a^{(k)}}\right]\right)dy.$$

By using the dominated convergence theorem, we have $\lim_{h\to 0}\frac{\epsilon_3(\nu,h)}{h}=0$ for all ν.

Finally, by using the fact that f is a bounded function, we now prove that the fourth term on the right side of equation (50) is a $o(h)$ term denoted by $\epsilon_4(\nu,h)$. Since $f\in\mathcal{C}_b(\mathcal{M}_F^N(\mathcal{U}))$, it is enough to prove that $\sum_{i\geq 1,j\geq 1}\mathbb{P}(\{A_h=i,D_h=j\}|\eta_0^N=\nu)$ is a $o(h)$ term. In this direction, we specify that $\left(\sum_{i\geq 2,j\geq 1}\mathbb{P}(\{A_h=i,D_h=j\}|\eta_0^N=\nu)\right)\leq\mathbb{P}(\{A_h\geq 2\})$. Since $\mathbb{P}(\{A_h\geq 2\}$ is a $o(h)$ term, we have that $\sum_{i\geq 2,j\geq 1}\mathbb{P}(\{A_h=i,D_h=j\}|\eta_0^N=\nu)$ is a $o(h)$ term for all ν. We now show that $\sum_{j\geq 1}\mathbb{P}(\{A_h=1,D_h=j\}|\eta_0^N=\nu)$ is a $o(h)$ term. We can write

$$\Big(\sum_{j\geq 1}\mathbb{P}(\{A_h=1,D_h=j\}|\eta_0^N=\nu)\Big)=\mathbb{P}(\{A_h=1\}|\eta_0^N=\nu)-\mathbb{P}(\{A_h=1,D_h=0\}|\eta_0^N=\nu)$$
$$=\mathbb{P}(\{A_h=1\})(1-\mathbb{P}(\{D_h=0\}|A_h=1,\eta_0^N=\nu)).$$

Again, by using the method of change of variables and the dominated convergence theorem as in the proof of the result that $\epsilon_3(\nu,h)$ is a $o(h)$ term, we get that $\lim_{h\to 0}\mathbb{P}(\{D_h=0\}|A_h=$

$1, \eta_0^N = \nu) = 1$ for all ν. Since $\lim_{h\to 0} \frac{\mathbb{P}(\{A_h=1\})}{h} = N\lambda$, we have that $\left(\sum_{j\geq 1} \mathbb{P}(\{A_h = 1, D_h = j\}|\eta_0^N = \nu)\right)$ is a $o(h)$ term for all ν. Therefore, $\epsilon_4(\nu, h)$ is a $o(h)$ term for all ν.

By combining the expressions for all the four terms on the right side of equation (50), and by defining $\epsilon(\nu, h) = \epsilon_1(\nu, h) + \epsilon_2(\nu, h) + \epsilon_3(\nu, h) + \epsilon_4(\nu, h)$, we get the expression for $T_h^N f(\nu)$ as in equation (49). Finally, from [10, p.18], $(\eta_t^N, t \geq 0)$ is a weak homogeneous Markov process.

Finally, the proof of Feller-Dynkin property follows *mutatis mutandis* from the proof of Proposition 1 of [12]. □

Appendix B. Proof of Lemma 5.1

Proof. We first show that any process $(\nu_t, t \geq 0)$ that satisfies equation (9) also satisfies equation (10). By using the fundamental theorem of calculus, for $\phi \in \mathcal{C}_b^1(\mathcal{U})$, a real valued process $(\langle \nu_t, \phi\rangle, t \geq 0)$ satisfying equation (9) is a solution to the following differential equation (52) if the integrand in equation (9) is a continuous function of s,

$$\frac{d\langle \nu_t, \phi\rangle}{dt} = \langle \nu_t, \nabla_1 \phi\rangle + \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right) d\nu_t(\mathbf{x}_n) \right.$$
$$+ \left[\nu_t(\{0\}) \lambda \Phi_0(\nu_t) \left(\phi(1,0) - \phi(0)\right) + \sum_{n=1}^{C-1} \sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \right.$$
$$\left. \left. \times \lambda \Phi_n(\nu_t) (\phi(\mathbf{x}_n^j; 0) - \phi(\mathbf{x}_n)) \, d\nu_t(\mathbf{x}_n) \right] \right). \quad (52)$$

Therefore we need to show that the two terms on the right side of equation (52) are continuous functions of t. Since $\phi \in \mathcal{C}_b^1(\mathcal{U})$ and the mapping $t \mapsto \nu_t$ is continuous, the first term $\langle \nu_t, \nabla_1 \phi\rangle$ is a continuous function of t. In the second term, the expression related to the case of departures can be written as

$$\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right) d\nu_t(\mathbf{x}_n) = \langle \nu_t, \psi_1\rangle,$$

where the function ψ_1 is defined as

$$\psi_1(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = 0, \\ \sum_{j=1}^{n} \beta(x_j)((\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)) & \text{otherwise.} \end{cases}$$

Since $\phi \in \mathcal{C}_b^1(\mathcal{U})$ and $\beta \in \mathcal{C}_b^1(\mathbb{R}_+)$, we have that $\psi_1 \in \mathcal{C}_b(\mathcal{U})$. Therefore the mapping $t \mapsto \langle \nu_t, \psi_1 \rangle$ is continuous. The expression that corresponds to the case of arrivals can be written as

$$\begin{aligned}\langle \nu_t, \psi_{(\nu_t)} \rangle = (\nu_t(\{0\})\lambda\Phi_0(\nu_t)\left(\phi(1,0) - \phi(0)\right)) \\ + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \lambda\Phi_n(\nu_t)(\phi(\mathbf{x}_n^j;0) - \phi(\mathbf{x}_n))\, d\nu_t(\mathbf{x}_n),\end{aligned}$$

where $\psi_{(\nu_t)}$ is defined as

$$\psi_{(\nu_t)}(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = C, \\ \frac{\lambda\Phi_n(\nu_t)}{(n+1)} \sum_{j=1}^{n+1} (\phi(\mathbf{x}_n^j;0) - \phi(\mathbf{x}_n)) & \text{otherwise.} \end{cases} \tag{53}$$

For given ν_t, since $\phi \in \mathcal{C}_b(\mathcal{U})$, we have that $\psi_{(\nu_t)} \in \mathcal{C}_b(\mathcal{U})$. Hence for any constant $a \geq 0$, the mapping $t \mapsto \langle \nu_t, \psi_{(\nu_a)} \rangle$ is continuous.

We next prove that the mapping $t \mapsto \langle \nu_t, \psi_{(\nu_t)} \rangle$ is continuous, $i.e.$, we need to prove that $\langle \nu_{t+b}, \psi_{(\nu_{t+b})} \rangle \to \langle \nu_t, \psi_{(\nu_t)} \rangle$ as $b \to 0$. We have

$$\left|\langle \nu_{t+b}, \psi_{(\nu_{t+b})} \rangle - \langle \nu_t, \psi_{(\nu_t)} \rangle\right| \leq \left|\langle \nu_{t+b}, \psi_{(\nu_{t+b})} \rangle - \langle \nu_{t+b}, \psi_{(\nu_t)} \rangle\right| + \left|\langle \nu_{t+b}, \psi_{(\nu_t)} \rangle - \langle \nu_t, \psi_{(\nu_t)} \rangle\right|. \tag{54}$$

Since $\psi_{(\nu_t)} \in \mathcal{C}_b(\mathcal{U})$, we have that $\lim_{b\to 0} \left|\langle \nu_{t+b}, \psi_{(\nu_t)} \rangle - \langle \nu_t, \psi_{(\nu_t)} \rangle\right| = 0$. We next prove that $\lim_{b\to 0} \left|\langle \nu_{t+b}, \psi_{(\nu_{t+b})} - \psi_{(\nu_t)} \rangle\right| = 0$.

For $L > 0$, let

$$U^{(L)} = \{\mathbf{x}_n \in \mathcal{U}_n : n \geq 1, x_i > L \text{ for all } 1 \leq i \leq n\}.$$

For given $\epsilon > 0$, since ν_t is tight, we can find some $L > 0$ such that $\langle \nu_t, I_{\left\{U^{(L)}\right\}} \rangle < \epsilon$. Furthermore, from the continuity of the mapping $t \mapsto \nu_t$, we can find some $h_1 > 0$ such that for all $b \in [-\min(t, h_1), h_1]$,

$$\langle \nu_{t+b}, I_{\left\{U^{(L)}\right\}} \rangle < \epsilon. \tag{55}$$

By using the fact that the mapping $t \mapsto \overline{R}_n(\nu_t) = \langle \nu_t, I_{\left\{\cup_{j=n}^{C} \mathcal{U}_j\right\}} \rangle$ is continuous, we next show that the mapping $t \mapsto \psi_{(\nu_t)}$ is continuous. For this, we need to show that $\|\psi_{(\nu_{t+b})} - \psi_{(\nu_t)}\| \to 0$ as $b \to 0$. From equation (53), we have

$$\begin{aligned}\|\psi_{(\nu_{t+b})} - \psi_{(\nu_t)}\| &\leq 2\lambda\|\phi\| \max_n(|\Phi_n(\nu_{t+b}) - \Phi_{n+1}(\nu_t)|) \\ &\leq 4d\lambda\|\phi\| \max_n(\left|\overline{R}_n(\nu_{t+b}) - \overline{R}_n(\nu_t)\right|).\end{aligned} \tag{56}$$

Since $\left|\overline{R}_n(\nu_{t+b}) - \overline{R}_n(\nu_t)\right| \to 0$ as $b \to 0$ for all n, $\|\psi_{(\nu_{t+b})} - \psi_{(\nu_t)}\| \to 0$. This proves that the mapping $t \mapsto \psi_{(\nu_t)}$ is continuous. As a consequence, we have that $\psi_{(\nu_{t+b})}$ is uniformly continuous on the interval $b \in [-\min(t, h_1), h_1]$ and $\mathbf{u} \in \overline{U}^{(L)}$ (the complement of $U^{(L)}$). As a result, there exists some $h_2 \in (0, h_1)$ such that for $b \in [-\min(t, h_2), h_2]$, $\mathbf{u} \in \overline{U}^{(L)}$, we have

$$\left|\psi_{(\nu_{t+b})}(\mathbf{u}) - \psi_{(\nu_t)}(\mathbf{u})\right| < \epsilon. \tag{57}$$

Using equations (55)-(57), for $b \in [-\min(t, h_2), h_2]$, we have

$$\left|\langle \nu_{t+b}, \psi_{(\nu_{t+b})} - \psi_{(\nu_t)} \rangle\right| \leq \epsilon \langle \nu_{t+b}, I_{\left\{\overline{U}^{(L)}\right\}} \rangle + 4d\lambda\|\phi\|\epsilon \leq \epsilon + 4d\lambda\|\phi\|\epsilon. \tag{58}$$

By letting $b \to 0$ and then $\epsilon \to 0$ in equation (54), we have the continuity of the mapping $t \mapsto \langle \nu_t, \psi_{(\nu_t)} \rangle$.

We next show that a solution of equation (52) is also a solution to another differential equation obtained by applying a method of change of variables. For $r \leq t$, we have

$$\begin{aligned}\frac{d\langle \nu_r, \tau_{t-r}\phi\rangle}{dr} &= \lim_{h\to 0} \frac{(\langle \nu_{r+h}, \tau_{t-r-h}\phi\rangle - \langle \nu_r, \tau_{t-r}\phi\rangle)}{h} \\ &= \lim_{h\to 0} \frac{(\langle \nu_{r+h}, \tau_{t-r-h}\phi\rangle - \langle \nu_{r+h}, \tau_{t-r}\phi\rangle)}{h} + \lim_{h\to 0} \frac{(\langle \nu_{r+h}, \tau_{t-r}\phi\rangle - \langle \nu_r, \tau_{t-r}\phi\rangle)}{h}\end{aligned} \tag{59}$$

We now look at the first term on the right side of equation (59). We can write

$$\langle \nu_{r+h}, \tau_{t-r-h}\phi\rangle - \langle \nu_{r+h}, \tau_{t-r}\phi\rangle = \langle \nu_{r+h}, \hat{w}\rangle,$$

where $\hat{w}$ is defined such that $\hat{w}(\mathbf{y}_n) = \tau_{t-r-h}\phi(\mathbf{y}_n) - \tau_{t-r}\phi(\mathbf{y}_n)$. We further simplify the function $\hat{w}$ by using the following definition, let

$$\frac{\partial \phi}{\partial s_i}(\mathbf{y}_n) = \lim_{h\to 0} \frac{\phi(\mathbf{y}_n^{-j}; y_i + h) - \phi(\mathbf{y}_n)}{h}.$$

We can write

$$\hat{w}(\mathbf{y}_n) = \phi(\tau^+_{t-r-h}(\mathbf{y}_n)) - \phi((\tau^+_{t-r-h}(\mathbf{y}_n))^{-1}; y_1 + t - r) + \phi((\tau^+_{t-r-h}(\mathbf{y}_n))^{-1}; y_1 + t - r) - \phi(\tau^+_{t-r}(\mathbf{y}_n))$$

Further, we have

$$\phi(\tau^+_{t-r-h}(\mathbf{y}_n)) - \phi((\tau^+_{t-r-h}(\mathbf{y}_n))^{-1}; y_1 + t - r) = -\int_{y_1+t-r-h}^{y_1+t-r} \frac{\partial \phi}{\partial s_1}((\tau^+_{t-r-h}(\mathbf{y}_n))^{-1}; s_1)\, ds_1.$$

By replacing s_1 with $y_1 + t - r - hv$, we get

$$\phi(\tau^+_{t-r-h}(\mathbf{y}_n)) - \phi((\tau^+_{t-r-h}(\mathbf{y}_n))^{-1}; y_1 + t - r) = -h\int_{v=0}^{1} \frac{\partial \phi}{\partial s_1}((\tau^+_{t-r-h}(\mathbf{y}_n))^{-1}; y_1 + t - r - hv)\, dv.$$

Similarly, we can write

$$\begin{aligned}
&\phi(n, y_1 + t - r, \cdots, y_{i-1} + t - r, y_i + t - r - h, y_{i+1} + t - r - h, \cdots, y_n + t - r - h) \\
&\qquad - \phi(n, y_1 + t - r, \cdots, y_i + t - r, y_{i+1} + t - r - h, \cdots, y_n + t - r - h) \\
&= -h \int_{v=0}^{1} \frac{\partial \phi}{\partial s_i}(n, y_1 + t - r, \cdots, y_{i-1} + t - r, y_i + t - r - hv, y_{i+1} + t - r - h, \cdots, y_n + t - r - h)\, dv.
\end{aligned}$$

For $1 \leq i \leq n$, let

$$w_{(i,t,r,h,v)}(\mathbf{y}_n) = \frac{\partial \phi}{\partial s_i}(n, y_1 + t - r, \cdots, y_{i-1} + t - r, y_i + t - r - hv, y_{i+1} + t - r - h, \cdots, y_n + t - r - h)$$

As a consequence, after simplifications, we have

$$\hat{w}(\mathbf{y}_n) = -h \int_{v=0}^{1} \sum_{i=1}^{n} \left(w_{(i,t,r,h,v)}(\mathbf{y}_n) \right) \, dv$$

Let the function $w^*_{(t,r,h,v)} \in \mathcal{C}_b(\mathcal{U})$ be defined as

$$w^*_{(t,r,h,v)}(\mathbf{y}_n) = \begin{cases} 0 & \text{if } n = 0, \\ \sum_{i=1}^{n} \left(w_{(i,t,r,h,v)}(\mathbf{y}_n) \right) & \text{otherwise.} \end{cases}$$

Now we can see that

$$\lim_{h \to 0} \frac{\left(\langle \nu_{r+h}, \tau_{t-r-h} \phi \rangle - \langle \nu_{r+h}, \tau_{t-r} \phi \rangle \right)}{h} = - \lim_{h \to 0} \int_{v=0}^{1} \langle \nu_{r+h}, w^*_{(t,r,h,v)} \rangle \, dv.$$

Since $h \mapsto \langle \nu_{r+h}, w^*_{(t,r,h,v)} \rangle$ is continuous, by the dominated convergence theorem, we have

$$\lim_{h \to 0} \frac{\left(\langle \nu_{r+h}, \tau_{t-r-h} \phi \rangle - \langle \nu_{r+h}, \tau_{t-r} \phi \rangle \right)}{h} = -\langle \nu_r, \nabla_1 \tilde{\phi} \rangle. \tag{60}$$

We now look at the second term on the right side of equation (59). We have

$$\langle \nu_{r+h}, \tau_{t-r} \phi \rangle - \langle \nu_r, \tau_{t-r} \phi \rangle = \int_{u=r}^{r+h} \frac{\partial}{\partial u} \langle \nu_u, \tau_{t-r} \phi \rangle \, du$$

By using equation (52), we have

$$\begin{aligned}
\langle \nu_{r+h}, \tau_{t-r} \phi \rangle - \langle \nu_r, \tau_{t-r} \phi \rangle = \int_{u=r}^{r+h} \Bigg(& \langle \nu_u, \nabla_1 \tau_{t-r} \phi \rangle \\
& + \sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\tau_{t-r} \phi(\mathbf{x}_n^{-j}) - \tau_{t-r} \phi(\mathbf{x}_n) \right) \, d\nu_u(\mathbf{x}_n) \\
& + \Bigg[\nu_u(\{0\}) \lambda \Phi_0(\nu_u) \left(\tau_{t-r} \phi(1, 0) - \tau_{t-r} \phi(0) \right) + \sum_{n=1}^{C-1} \sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{(n+1)} \\
& \qquad \times \lambda \Phi_n(\nu_u) (\tau_{t-r} \phi(\mathbf{x}_n^j; 0) - \tau_{t-r} \phi(\mathbf{x}_n)) \, d\nu_u(\mathbf{x}_n) \Bigg] \Bigg) \, du.
\end{aligned}$$

Again, by using change of variables, we have

$$\langle \nu_{r+h}, \tau_{t-r}\phi\rangle - \langle \nu_r, \tau_{t-r}\phi\rangle = h\int_{v=0}^{1} \langle \nu_{r+hv}, \nabla_1 \tau_{t-r}\phi\rangle + \left(\sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j)\left(\tau_{t-r}\phi(\mathbf{x}_n^{-j}) - \tau_{t-r}\phi(\mathbf{x}_n)\right) d\nu_{r+hv}(\mathbf{x}_n) + \left[\nu_{r+hv}(\{0\})\lambda\Phi_0(\nu_{r+hv})\left(\tau_{t-r}\phi(1,0) - \tau_{t-r}\phi(0)\right) + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1}\int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)} \times \lambda\Phi_n(\nu_{r+hv})(\tau_{t-r}\phi(\mathbf{x}_n^j;0) - \tau_{t-r}\phi(\mathbf{x}_n))\, d\nu_{r+hv}(\mathbf{x}_n)\right]\right) dv.$$

As a result, by using the dominated convergence theorem, we have

$$\lim_{h\to 0}\frac{\nu_{r+h}, \tau_{t-r}\phi\rangle - \langle \nu_r, \tau_{t-r}\phi\rangle}{h} = \langle \nu_r, \nabla_1 \tau_{t-r}\phi\rangle + \left(\sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j)\left(\tau_{t-r}\phi(\mathbf{x}_n^{-j}) - \tau_{t-r}\phi(\mathbf{x}_n)\right) d\nu_r(\mathbf{x}_n) + \left[\nu_r(\{0\})\lambda\Phi_0(\nu_r)\left(\tau_{t-r}\phi(1,0) - \tau_{t-r}\phi(0)\right) + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1}\int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)} \times \lambda\Phi_n(\nu_r)(\tau_{t-r}\phi(\mathbf{x}_n^j;0) - \tau_{t-r}\phi(\mathbf{x}_n))\, d\nu_r(\mathbf{x}_n)\right]\right) \quad (61)$$

Finally, by using equations (60) and (61), we have

$$\frac{d\langle \nu_r, \tau_{t-r}\phi\rangle}{dr} = \sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j)\left(\tau_{t-r}\phi(\mathbf{x}_n^{-j}) - \tau_{t-r}\phi(\mathbf{x}_n)\right) d\nu_r(\mathbf{x}_n) + \left[\nu_r(\{0\})\lambda\Phi_0(\nu_r)\left(\tau_{t-r}\phi(1,0) - \tau_{t-r}\phi(0)\right) + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1}\int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)} \times \lambda\Phi_n(\nu_r)(\tau_{t-r}\phi(\mathbf{x}_n^j;0) - \tau_{t-r}\phi(\mathbf{x}_n))\, d\nu_r(\mathbf{x}_n)\right].$$

By integrating $\frac{d\langle \nu_r, \tau_{t-r}\phi\rangle}{dr}$ with respect to r from 0 to t, we get equation (10) for $\phi \in \mathcal{C}_b^1(\mathcal{U})$. Then the result can be extended to the simple functions by using the monotone convergence theorem and then to the class of functions $\mathcal{C}_b(\mathcal{U})$ from the standard arguments by using the Dynkin $\pi - \lambda$ theorem [13, page 497].

We next prove that for $\phi \in \mathcal{C}_b^1(\mathcal{U})$, the solution $(\langle \overline{\eta}_t, \phi\rangle, t \geq 0)$ of equation (10) is a solution to equation (9). For this, it is enough to prove the differentiability of $\langle \overline{\eta}_t, \phi\rangle$ with respect to t. Since $\phi \in \mathcal{C}_b^1(\mathcal{U})$, the existence of $\frac{d\langle \overline{\eta}_0, \tau_t\phi\rangle}{dt}$ follows from the dominated convergence

theorem. By using the Leibniz integral rule, we verify the existence of the differentiation of the second term on the right side of equation (10) with respect to t. According to this rule, the first condition is that the integrand needs to be continuous with respect to both the variables r and t. This follows from the same arguments that we have used to prove the continuity of the integrand in equation (9). The second condition is that the differentiation of the integrand with respect to t must exist and the differential should be continuous with respect to both r and t. The differential of the integrand with respect to t exists from the dominated convergence theorem since $\phi \in \mathcal{C}_b^1(\mathcal{U})$ and also, it is continuous with respect to r and t from the same arguments that we have used to prove the continuity of the integrand in equation (9). Therefore any process $(\nu_t, t \geq 0) \in \mathcal{C}_{\mathcal{M}_1(\mathcal{U})}([0, \infty))$ is a solution to equation (9) if and only if it is solution to equation (10). Further, note that ϕ need not be a differentiable function in equation (10). □

Appendix C. Proof of Theorem 5.1

Proof. From equation (10), we first make it clear that for all $\phi \in \mathcal{C}_b(\mathcal{U})$, the operator $\phi \mapsto \langle \nu_t, \phi \rangle$ is a linear operator with $\nu_t(\mathcal{U}) = 1$. Hence from the Riesz-Markov-Kakutani theorem [40, Theorem 2.14], for $\nu_t \in \mathcal{M}_1(\mathcal{U})$, the existence of the unique operator $\phi \mapsto \langle \nu_t, \phi \rangle$ implies the existence of the unique probability measure ν_t. The uniqueness of ν_t also follows from the fact that $\mathcal{C}_b(\mathcal{U})$ is a separating class of $\mathcal{M}_1(\mathcal{U})$ [13, p.111], if $\eta_1, \eta_2 \in \mathcal{M}_1(\mathcal{U})$ satisfies $\langle \nu_t, \phi \rangle = \langle \eta_1, \phi \rangle$ and $\langle \nu_t, \phi \rangle = \langle \eta_2, \phi \rangle$ for all $\phi \in \mathcal{C}_b(\mathcal{U})$, then we have $\eta_1 = \eta_2$.

Given an initial measure ν_0, we next prove that there exists at most one mean-field solution by showing that there exists at most one real valued process $\langle \nu_t, \phi \rangle$ corresponding to the mean-field. Suppose $(\nu_t^1, t \geq 0), (\nu_t^2, t \geq 0)$ are two solutions to the mean-field equations with initial

points ν_0^1, ν_0^2, respectively. For $\phi \in \mathcal{C}_b(\mathcal{U})$, we then have

$$\begin{aligned}\langle \nu_t^1 - \nu_t^2, \phi\rangle = \langle \nu_0^1 - \nu_0^2, \tau_t\phi\rangle + \int_{s=0}^{t} \Bigg(\sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j)\left(\tau_{t-s}\phi(\mathbf{x}_n^{-j}) - \tau_{t-s}\phi(\mathbf{x}_n)\right) \\ \times d(\nu_s^1 - \nu_s^2)(\mathbf{x}_n)\Bigg) ds \\ + \int_{s=0}^{t} \Bigg(\Bigg[\nu_s^1(\{0\})\lambda\Phi_0(\nu_s^1)\left(\tau_{t-s}\phi(1,0) - \tau_{t-s}\phi(0)\right) \\ + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1} \int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)} \lambda\Phi_n(\nu_s^1)(\tau_{t-s}\phi(\mathbf{x}_n^j;0) - \tau_{t-s}\phi(\mathbf{x}_n))\, d\nu_s^1(\mathbf{x}_n)\Bigg] \\ - \Bigg[\nu_s^2(\{0\})\lambda\Phi_0(\nu_s^2)\left(\tau_{t-s}\phi(1,0) - \tau_{t-s}\phi(0)\right) \\ - \sum_{n=1}^{C-1}\sum_{j=1}^{n+1} \int\cdots\int_{\mathcal{U}_n} \frac{1}{(n+1)} \lambda\Phi_n(\nu_s^2)(\tau_{t-s}\phi(\mathbf{x}_n^j;0) - \tau_{t-s}\phi(\mathbf{x}_n))\, d\nu_s^2(\mathbf{x}_n)\Bigg]\Bigg) ds. \quad (62)\end{aligned}$$

The first term on the right side of equation (62) can be bounded as $\left|\langle \nu_0^1 - \nu_0^2, \tau_t\phi\rangle\right| \leq \|\nu_0^1 - \nu_0^2\|\|\phi\|$. To simplify the second term corresponding to departures, we define a function $h_{t,s}$ as follows:

$$h_{t,s}(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = 0, \\ \sum_{k=1}^{n} \beta(x_k)(\tau_{t-s}\phi(\mathbf{x}_n^{-j}) - \tau_{t-s}\phi(\mathbf{x}_n)) & \text{otherwise.}\end{cases}$$

Then since $\phi \in \mathcal{C}_b(\mathcal{U})$ and $\beta \in \mathcal{C}_b(\mathbb{R}_+)$, we have $h_{t,s} \in \mathcal{C}_b(\mathcal{U})$. Further, we have $\|h_{t,s}\| \leq 2C\|\beta\|\|\phi\|$. Using the definition of $h_{t,s}$, we have

$$\begin{aligned}\int_{s=0}^{t} \Bigg(\sum_{n=1}^{C}\sum_{j=1}^{n} \int\cdots\int_{\mathcal{U}_n} \beta(x_j)\Bigg(\tau_{t-s}\phi(\mathbf{x}_n^{-j}) - \tau_{t-s}\phi(\mathbf{x}_n)\Bigg) d(\nu_s^1 - \nu_s^2)(\mathbf{x}_n)\, ds \\ = \int_{s=0}^{t} \langle \nu_s^1 - \nu_s^2, h_{t,s}\rangle ds.\end{aligned}$$

To simplify the third term corresponding to arrivals, we define a function $f_{t,s,\nu}$ as follows: for $0 \leq n \leq C-1$,

$$f_{t,s,\nu}(\mathbf{x}_n) = \begin{cases} 0 & \text{if } n = C, \\ \sum_{j=1}^{n+1} \frac{1}{(n+1)} \Phi_n(\nu)(\tau_{t-s}\phi(\mathbf{x}_n^j;0) - \tau_{t-s}\phi(\mathbf{x}_n)) & \text{otherwise.}\end{cases}$$

Then the third term is equal to $\int_{s=0}^{t} \lambda \left(\langle \nu_s^1, f_{t,s,\nu_s^1} \rangle - \langle \nu_s^2, f_{t,s,\nu_s^2} \rangle \right) \, ds$. Further, we can write

$$\begin{aligned} \left| \langle \nu_s^1, f_{t,s,\nu_s^1} \rangle - \langle \nu_s^2, f_{t,s,\nu_s^2} \rangle \right| &\leq \left| \langle \nu_s^1 - \nu_s^2, f_{t,s,\nu_s^1} \rangle \right| + \left| \langle \nu_s^2, f_{t,s,\nu_s^1} - f_{t,s,\nu_s^2} \rangle \right| \\ &\leq \|\nu_s^1 - \nu_s^2\| \|f_{t,s,\nu_s^1}\| + \|\nu_s^2\| \|f_{t,s,\nu_s^1} - f_{t,s,\nu_s^1}\|. \end{aligned}$$

Since ν_s^2 is a probability measure, $\|\nu_s^2\| = 1$. Furthermore, $\|f_{t,s,\nu_s^1}\| \leq 2d\|\phi\|$ and

$$\left| f_{t,s,\nu_s^1}(\mathbf{x}_n) - f_{t,s,\nu_s^2}(\mathbf{x}_n) \right| \leq 2d^2 \|\phi\| \left(\left| \overline{R}_n(\nu_s^1) - \overline{R}_n(\nu_s^2) \right| + \left| \overline{R}_{n+1}(\nu_s^1) - \overline{R}_{n+1}(\nu_s^2) \right| \right).$$

We can write $\overline{R}_n(\nu_s^1) = \langle \nu_s^1, f^* \rangle$ where f^* is a function defined as

$$f^*(\mathbf{x}_m) = \begin{cases} 1 & \text{if } m \geq n, \\ 0 & \text{otherwise.} \end{cases}$$

We then have $\left| \overline{R}_n(\nu_s^1) - \overline{R}_n(\nu_s^2) \right| \leq \|\nu_s^1 - \nu_s^2\| \|f^*\| = \|\nu_s^1 - \nu_s^2\|$.

Finally, by using bounds for all the terms, we get

$$\left| \langle \nu_t^1 - \nu_t^2, \phi \rangle \right| \leq \left(\|\nu_0^1 - \nu_0^2\| + \int_{s=0}^{t} 2\|\beta\| C \|\nu_s^1 - \nu_s^2\| \, ds \; + \int_{s=0}^{t} 8d^2 \lambda \|\nu_s^1 - \nu_s^2\| \, ds \right) \|\phi\|.$$

Therefore we have

$$\|\nu_t^1 - \nu_t^2\| \leq \|\nu_0^1 - \nu_0^2\| + (2C\|\beta\| + 8d^2\lambda) \int_{s=0}^{t} \|\nu_s^1 - \nu_s^2\| \, ds. \tag{63}$$

From the Gronewall's inequality, for some $b, c > 0, t \in [0, T]$, if $\|\nu_t^1 - \nu_t^2\| \leq b + c \int_{s=0}^{t} \|\nu_s^1 - \nu_s^2\| ds$, then it follows that $\|\nu_t^1 - \nu_t^2\| \leq b \, e^{ct}$. Therefore, from equation (63), we have $\|\nu_t^1 - \nu_t^2\| \leq \|\nu_0^1 - \nu_0^2\| \, e^{(2C\|\beta\| + 8d^2\lambda)t}$. Hence, starting from an initial measure ν_0, there exists at most one solution for the mean-field equations.

We now prove that there exists a process $(\nu_t, t \geq 0) \in \mathcal{C}_{\mathcal{M}_1(\mathcal{U})}([0, \infty))$ satisfying the mean-field model equations. This follows from the relative compactness of the sequence $\{\overline{\eta}_t^N, t \geq 0\}$ in $\mathcal{D}_{\mathcal{M}_1(\mathcal{U})}([0, \infty))$ from the proof of Theorem 5.2. In particular, we have that every limit point of the sequence $\{\overline{\eta}_t^N, t \geq 0\}$ satisfies equation (10). Further, each limiting point is almost surely continuous. This concludes that there exists a solution to the mean-field equations.

□

Appendix D. Martingale construction

In this section, by using the infinitesimal generator of the Markov process $(\eta_t^N, t \geq 0)$, we construct a martingale $(M_t^N(\phi), t \geq 0) \in D_{\mathbb{R}}([0, \infty))$ where $\phi \in \mathcal{C}_b^1(\mathcal{U})$. We then show

that the scaled version of the process $(M_t^N(\phi), t \geq 0)$ converges in distribution to the null process based on which we later establish the convergence of the scaled version of the process $(\eta_t^N, t \geq 0)$.

Since the set of linear combinations of $Q_f : \mathcal{M}_F(\mathcal{U}) \longmapsto \mathbb{R}$ for $f \in \mathcal{C}_s^1(\mathcal{U})$ defined by $Q_f(\nu) = e^{-\langle \nu, f\rangle}$ is dense in the set $\mathcal{C}(\mathcal{M}_F(\mathcal{U}))$ [38, proposition 7.10], by using $A^N Q_f(\nu)$, for any continuous function $F \in \mathcal{C}(\mathcal{M}_F(\mathcal{U}))$ such that the infinitesimal generator $A^N F(\nu) = \lim_{h\to 0} \frac{\mathbb{E}[F(\eta_h^N)|\eta_0^N=\nu]-F(\nu)}{h}$ is well-defined, we have for all ν

$$A^N F(\nu) = \lim_{h\to 0} \frac{F(\tau_h \nu) - F(\nu)}{h} - N\lambda F(\nu) - F(\nu) \sum_{n=1}^{C}\sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j)\, d\nu(\mathbf{x}_n)$$
$$+ \sum_{n=1}^{C}\sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(F(\nu + \delta_{(\mathbf{x}_n^{-j})} - \delta_{(\mathbf{x}_n)}\right) d\nu(\mathbf{x}_n)$$
$$+ N\lambda \Bigg[\left(\frac{\nu(\{0\})}{N} \Phi_0 \left(\frac{\nu}{N}\right) \left(F(\nu + \delta_{(1,0)} - \delta_{(0)})\right)\right) + \sum_{n=1}^{C-1}\sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{N(n+1)} \Phi_n \left(\frac{\nu}{N}\right)$$
$$\times F(\nu + \delta_{(\mathbf{x}_n^j;0)} - \delta_{(\mathbf{x}_n)})\, d\nu(\mathbf{x}_n) + \int \cdots \int_{\mathcal{U}_C} \frac{1}{N} \Phi_C \left(\frac{\nu}{N}\right) F(\nu)\, d\nu(\mathbf{x}_C) \Bigg]. \quad (64)$$

For $\phi \in \mathcal{C}_b^1(\mathcal{U})$, it is clear that the function $\nu \in \mathcal{M}_F(\mathcal{U}) \mapsto \langle \nu, \phi\rangle \in \mathbb{R}$ belongs to the domain of A^N.

Proposition D.1. *For all $\phi \in \mathcal{C}_b^1(\mathcal{U})$, the process $(M_t^N(\phi), t \geq 0)$ given by*

$$M_t^N(\phi) = \langle \eta_t^N, \phi\rangle - \langle \eta_0^N, \phi\rangle - \int_{s=0}^{t} A^N \langle \eta_s^N, \phi\rangle ds \quad (65)$$

is a RCLL (process that is right continuous with left limits) square integrable $\mathcal{F}_t^N-$martingale. For $\phi \in \mathcal{C}_b^1(\mathcal{U})$, the quadratic variation of $(M_t^N(\phi), t \geq 0)$ is given by

$$< M_\cdot^N(\phi) >_t = \int_{s=0}^{t} \Bigg(\sum_{n=1}^{C}\sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right)^2 d\eta_s^N(\mathbf{x}_n)$$
$$+ N\lambda \Bigg[\left(\frac{\eta_s^N(\{0\})}{N} \Phi_0 \left(\frac{\eta_s^N}{N}\right) \left(\phi(1,0) - \phi(0)\right)^2 \right)$$
$$+ \sum_{n=1}^{C-1}\sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{N(n+1)} \Phi_n \left(\frac{\eta_s^N}{N}\right) \left(\phi(\mathbf{x}_n^j;0) - \phi(\mathbf{x}_n)\right)^2 d\eta_s^N(\mathbf{x}_n) \Bigg] \Bigg) ds \quad (66)$$

Proof. From the Dynkin's formula [13], the process $(M_t^N(\phi), t \geq 0)$ defined by

$$M_t^N(\phi) = \langle \eta_t^N, \phi\rangle - \langle \eta_0^N, \phi\rangle - \int_{s=0}^{t} A^N \langle \eta_s^N, \phi\rangle ds \quad (67)$$

is a RCLL $\mathcal{F}_t^N-$local martingale. Therefore, by simplification, we get

$$M_t^N(\phi) = \langle \eta_t^N, \phi\rangle - \langle \eta_0^N, \phi\rangle - \int_{s=0}^{t} \langle \eta_s^N, \nabla_1 \phi\rangle \, ds - \int_{s=0}^{t} \left(\sum_{n=1}^{C} \sum_{j=1}^{n} \int \cdots \int_{\mathcal{U}_n} \beta(x_j) \right.$$
$$\times \left(\phi(\mathbf{x}_n^{-j}) - \phi(\mathbf{x}_n)\right) d\eta_s^N(\mathbf{x}_n)$$
$$+ N\lambda \left[\left(\frac{\eta_s^N(\{0\})}{N} \Phi_0 \left(\frac{\eta_s^N}{N} \right) (\phi(1,0) - \phi(0)) \right) \right.$$
$$\left. \left. + \sum_{n=1}^{C-1} \sum_{j=1}^{n+1} \int \cdots \int_{\mathcal{U}_n} \frac{1}{N(n+1)} \Phi_n \left(\frac{\eta_s^N}{N} \right) (\phi(\mathbf{x}_n^j; 0) - \phi(\mathbf{x}_n)) \, d\eta_s^N(\mathbf{x}_n) \right] \right) ds. \quad (68)$$

By choosing $F_\phi(\eta_t^N) = \langle \eta_t^N, \phi\rangle$, from [11, Theorem 7.15], we have

$$< M_\cdot(\phi)^N >_t = \int_{s=0}^{t} \left(A^N F_\phi^2(\eta_s^N) - 2F_\phi(\eta_s^N) A^N F_\phi(\eta_s^N) \right) \, ds. \quad (69)$$

After simplifications, we get equation (66). Finally, since $\phi \in \mathcal{C}_b^1(\mathcal{U})$ and $\beta \in \mathcal{C}_b(\mathbb{R}_+)$, we have $\mathbb{E}\left[< M_\cdot^N(\phi) >_t\right] < \infty$ and hence, $(M_t^N(\phi), t \geq 0)$ is a square integrable martingale. □

Appendix E. Proof of Lemma 5.3:

Proof. Let us consider the function $\hat{\phi} = I_{\{\mathbf{l}_n \in \mathcal{U}_n : 0 \leq l_i \leq y_i, \, \forall i\}}$. For an absolutely continuous measure ν_s which has no atoms, we have $\langle \nu_s, \hat{\phi}\rangle = \langle \nu_s, \psi\rangle$, where $\psi = I_{\{\mathbf{u}_n \in \mathcal{U}_n : 0 < l_i < y_i, \, \forall i\}}$. Since there exists a sequence of functions $\{f_n\} \in \mathcal{C}_b(\mathcal{U})$ that increase point wise to $I_{\{B\}}$ where B is an open set in $\mathcal{U}_n$, $n \geq 1$, by using the monotone convergence theorem and equation (10), we have that equation (10) is true even for the function ψ (Indicators on open sets). Furthermore, since the measure ν_s is absolutely continuous for all $s \geq 0$, we have that equation (10) is true even for the function $\hat{\phi}$ (Indicators on closed sets). Therefore we can obtain the evolution equations for the process $(P_t, t \geq 0)$ that is defined as $P_t(\mathbf{y}_n) = \langle \nu_t, \hat{\phi}\rangle$ using equation (10). We can further simplify the expression of the process $(P_t(\mathbf{u}), \mathbf{u} \in \mathcal{U}, t \geq 0)$ obtained from equation (10) using the fact that

$$\langle \nu_s, \tau_b I_{\{\mathbf{x}_n \in \mathcal{U}_n : 0 \leq x_i \leq y_i, \, \forall i\}} \rangle = \langle \nu_s, I_{\{\mathbf{x}_n \in \mathcal{U}_n : 0 \leq x_i + b \leq y_i, \, \forall i\}} \rangle$$
$$= \langle \nu_s, I_{\{\mathbf{x}_n \in \mathcal{U}_n : 0 \leq x_i \leq y_i - b, \, \forall i\}} \rangle.$$

By differentiating $P_t(\mathbf{y}_n)$ with respect to t and after simplifications, it is verified that the process $P_t = (P_t(\mathbf{u}), \mathbf{u} \in \mathcal{U})$ satisfies equations (16)-(18).

□

Appendix F. Proof of Lemma 5.2

Proof. From the Remark 5.2, we recall that the MFEs are the dynamics of the probability distribution of a single server Loss system with capacity C where jobs arrive according to a Poisson process with rate $\lambda\Phi_n(\overline{\eta}_t)$ $(n \geq 1)$ when there are n progressing jobs. We have that the initial distribution ϑ has a density function and our objective is to show that for given $t = r$, $\overline{\eta}_r$ has a density function. For $n \geq 1$, $\mathbf{u}_n = (n, u_1, \cdots, u_n)$, we now prove that $\overline{\eta}_r$ has density at $\mathbf{u}_n$. For $\gamma_i > 0$, $1 \leq i \leq n$, let $B = ((n, y_1, \cdots, y_n) : u_i < y_i < u_i + \gamma_i, 1 \leq i \leq n)$. The probability that at time $t = r$, there are n progressing jobs and the i^{th} job has age y_i such that $y_i \in (u_i, u_i + \gamma_i)$, $i \geq 1$, is equal to $\overline{\eta}_r(B)$. Out of the n progressing jobs that are present at time $t = r$, let $\mathcal{J}_1$ be the set of indices of all the progressing jobs that entered the system at a time $t > 0$ and $\mathcal{J}_2$ be the set of indices of all the progressing jobs which are present from time $t = 0$. Precisely,

$$\mathcal{J}_1 = \{i : r \geq u_i, 1 \leq i \leq n\},$$

and

$$\mathcal{J}_2 = \{i : r < u_i, 1 \leq i \leq n\}.$$

Essentially, if $i \in \mathcal{J}_1$, it implies that the age of the i^{th} job is less than or equal to r and since the ages of progressing jobs increase linearly with time at unit rate, the i^{th} job must have entered the system at a time $t > 0$. Precisely, at time r, if the i^{th} job's age y_i satisfies $y_i \in (u_i, u_i + \gamma_i)$ and $i \in \mathcal{J}_1$, it implies that the i^{th} job must have entered the system in the time interval $(r - u_i - \gamma_i, r - u_i)$ and stayed in the system up to time $t = r$. On the other hand, if $j \in \mathcal{J}_2$, it implies that the j^{th} job is present in the system from time $t = 0$. At time $t = r$, if the j^{th} job's age y_j satisfies $y_j \in (u_j, u_j + \gamma_j)$ and $j \in \mathcal{J}_2$, then its age should lie in the interval $(u_j - r, u_j + \gamma_j - r)$ at time $t = 0$.

Using the sets $\mathcal{J}_1$ and $\mathcal{J}_2$, we now obtain an upper bound on $\overline{\eta}_r(B)$ from which we conclude that there exists a density function. For given set A, let $|A|$ be the number of elements in the set A. Further, let $J_1 = |\mathcal{J}_1|$ and $J_2 = |\mathcal{J}_2|$.

Let B_1 be the event that there exists at least J_2 jobs at time $t = 0$ such that for each $j \in \mathcal{J}_2$, there exists a job with age in the interval $(u_j - r, u_j + \gamma_j - r)$ and it should stay in the system up to time $t = r$. Note that the total number of jobs say q that are present at time $t = 0$ can be more than J_2, but only J_2 of them should stay in the system up to time $t = r$. A job with age x at time $t = 0$ will stay in the system at time $t = r$ with probability $\frac{\overline{G}(x+t)}{\overline{G}(x)}$. Let

$f_\vartheta = (f_\vartheta(\mathbf{u}), \mathbf{u} \in \mathcal{U})$ be the pdf of ϑ. Let l_i be the i^{th} smallest element of the set $\mathcal{J}_2$. Then by using all the above arguments, we get the following bound where q denotes the number of progressing jobs at time $t = 0$ and i_j is the index of the job out of q jobs which will stay in the system up to time $t = r$ with age lying in the interval $(u_{l_j}, u_{l_j} + \gamma_{l_j})$ at time $t = r$:

$$\mathbb{P}(B_1) \leq \sum_{q:q=J_2}^{C} \left(\sum_{(i_1,\cdots,i_{J_2})\in\{1,2,\cdots,q\}} \int \cdots \int_{\mathcal{V}} f_\vartheta(n, x_1, \cdots, x_n) \left(\prod_{m=1}^{J_2} \frac{\overline{G}(x_{i_m} + r)}{\overline{G}(x_{i_m})} \right) dx_1 \cdots dx_q \right), \tag{70}$$

where

$$\mathcal{V} = \{(x_1, \cdots, x_q) : x_m \in \mathbb{R}_+ \text{ if } m \notin \{i_1, \cdots, i_{J_2}\} \\ \text{and } x_m \in (u_{l_a} - r, u_{l_a} - r + \gamma_{l_a}) \text{ for } m = i_a, 1 \leq a \leq J_2, 1 \leq m \leq q\}.$$

We now focus on the jobs that belong to the set $\mathcal{J}_1$. Let B_2 be the event that for each $j \in \mathcal{J}_1$, there is an arrival in the time interval $(r - u_j - \gamma_j, r - u_j)$ and furthermore, this job should stay in the system until the time $t = r$. Since the arrival process is a Poisson process with rate $\lambda\Phi_n(\overline{\eta}_t)$ when there are n jobs and $\lambda\Phi_n(\overline{\eta}_t) \leq \lambda d$ for all $n \geq 0$, for any time interval $[t_1, t_2]$, we have

$$\mathbb{P}(X) \leq \mathbb{P}(Y),$$

where X denotes the number of arrivals to the server in the interval $[t_1, t_2]$ and Y denotes the number of arrivals in the interval $[t_1, t_2]$ when the arrival process is a Poisson process with rate λd. Let k_i be the i^{th} smallest element of the set $\mathcal{J}_1$. Then since the arrival instants have uniform distribution conditioned on the number of arrivals over a time interval [39, page 325], we get

$$\mathbb{P}(B_2) \leq \frac{(\lambda d)^{J_1}}{J_1!} \left(\prod_{j=1}^{J_1} \overline{G}(u_{k_j}) \gamma_{k_j} \right). \tag{71}$$

Finally, from (70) and (71), we have

$$\overline{\eta}_t(B) \leq \left(\sum_{q:q=J_2}^{C} \left(\sum_{(i_1,\cdots,i_{J_2})\in\{1,2,\cdots,q\}} \int \cdots \int_{\mathcal{V}} f_\vartheta(n, x_1, \cdots, x_n) \right. \right. \\ \left. \left. \left(\prod_{m=1}^{J_2} \frac{\overline{G}(x_{i_m} + r)}{\overline{G}(x_{i_m})} \right) dx_1 \cdots dx_q \right) \right) \left(\frac{(\lambda d)^{J_1}}{J_1!} \left(\prod_{j=1}^{J_1} \overline{G}(u_{k_j}) \gamma_{k_j} \right) \right). \tag{72}$$

Clearly, $\overline{\eta}_t$ has density at $\mathbf{u}$ since $\overline{\eta}_t(B) \to 0$ as $\gamma_j \to 0$ for $1 \leq j \leq n$. □

References

[1] AGHAJANI, R. AND RAMANAN, K. (2017). The hydrodynamic limit of a randomized load balancing network. *ArXiv e-prints*.

[2] AMAZON. Amazon EC2. `http://aws.amazon.com/ec2/`.

[3] ASMUSSEN, S. (2003). *Applied Probability and Queues* vol. 51 of *Stochastic Modelling and Applied Probability*. Springer, New York.

[4] BILLINGSLEY, P. (1999). *Convergence of probability measures* second ed. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

[5] BONALD, T. AND A.PROUTIERE (2002). Insensitivity in processor sharing networks. *Performance Evaluation* **49,** 193–209.

[6] BONALD, T., JONCKHEERE, M. AND PROUTIÉRE, A. (2004). Insensitive load balancing. In *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*. SIGMETRICS '04/Performance '04. ACM, New York, NY, USA. pp. 367–377.

[7] BRAMSON, M., LU, Y. AND PRABHAKAR, B. (2010). Randomized load balancing with general service time distributions. In *Proceedings of ACM SIGMETRICS*. pp. 275–286.

[8] BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S. AND ZHAO, L. (2005). Statistical analysis of a telephone call center. *Journal of the American Statistical Association* **100,** 36–50.

[9] BRUMELLE, S. L. (1978). A generalization of Erlang's loss system to state dependent arrival and service rates. *Mathematics of Operations Research* **3,** 10–16.

[10] DAWSON, D. A. (1993). *Measure-valued Markov processes* vol. 1541 of *École d'Été de Probabilités de Saint-Flour XXI—1991*. Springer, Berlin.

[11] DECREUSEFOND, L. AND M., P. (2012). *Stochastic Modeling and Analysis of Telecom Networks*. ISTE, London.

[12] DECREUSEFOND, L. AND MOYAL, P. (2008). A functional central limit theorem for the M/GI/∞ queue. *Ann. Appl. Probab.* **18,** 2156–2178.

[13] ETHIER, S. N. AND KURTZ, T. G. (1985). *Markov Processes: Characterization and Convergence*. John Wiley and Sons Ltd.

[14] GRAHAM, C. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability* **37,** 198–211.

[15] GRAHAM, C. (2005). Functional central limit theorems for a large network in which customers join the shortest of several queues. *Probability Theory and Related Fields* **131,** 97–120.

[16] GRAHAM, C. AND MÉLÉARD, S. (1993). Propagation of chaos for a fully connected loss network with alternate routing. *Stochastic Processes and their Applications* **44,** 159–180.

[17] GRAHAM, C. AND MÉLÉARD, S. (1997). Stochastic particle approximations for generalized boltzmann models and convergence estimates. *The Annals of Probability* **28,** 115–132.

[18] GROMOLL, H. C., PUHA, A. L. AND WILLIAMS, R. J. (2002). The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.* **12,** 797–859.

[19] GROMOLL, H. C., ROBERT, P. AND ZWART, B. (2008). Fluid limits for processor-sharing queues with impatience. *Math. Oper. Res.* **33,** 375–402.

[20] GÄRTNER, J. (1988). On the mckean-vlasov limit for interacting diffusions. *Mathematische Nachrichten* **137,** 197–248.

[21] JAKUBOWSKI, A. (1986). On the skorokhod topology. *Annales de l'I.H.P. Probabilités et Statistiques* **22,** 263–285.

[22] JAKUBOWSKI, A. (1986). On the skorokhod topology. *Annales de l'I.H.P. Probabilités et statistiques* **22,** 263–285.

[23] KALLENBERG, O. (1983). *Random measures*. Akademie-Verlag.

[24] KARPELEVICH, F. I. AND RYBKO, A. N. (2000). Thermodynamic limit for the mean field model of simple symmetrical closed queueing network. *Markov Processes and Related Fields* **6,** 89–105.

[25] KASPI, H. AND RAMANAN, K. (2011). Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* **21,** 33–114.

[26] KOLESAR, P. (1984). Stalking the endangered cat: A queueing analysis of congestion at automatic teller machines. *Interfaces* **14,** 16–26.

[27] KOLOKOLTSOV, V. N. (2010). *Nonlinear Markov Processes and Kinetic Equations*. Cambridge Tracts in Mathematics. Cambridge University Press.

[28] KOTELENEZ, P. M. AND KURTZ, T. G. (2008). Macroscopic limits for stochastic partial differential equations of mckean–vlasov type. *Probability Theory and Related Fields* **146,** 189.

[29] LI, Q.-L. AND LIN, C. (2006). The M/G/1 processor-sharing queue with disasters. *Computers & Mathematics with Applications* **51,** 987 – 998.

[30] MICROSOFT. Microsoft Azure. `http://www.microsoft.com/windowsazure/`.

[31] MITZENMACHER, M. (1996). The power of two choices in randomized load balancing. *PhD Thesis, Berkeley*.

[32] MUKHERJEE, D., BORST, S., VAN LEEUWAARDEN, J. AND WHITING, P. (2016). Universality of power-of-d load balancing schemes. *SIGMETRICS Perform. Eval. Rev.* **44,** 36–38.

[33] MUKHOPADHYAY, A. AND MAZUMDAR, R. R. (2014). Rate-based randomized routing in large heterogeneous processor sharing systems. In *Proceedings of 26th International Teletraffic Congress (ITC 26)*.

[34] MUKHOPADHYAY, A. AND MAZUMDAR, R. R. (2016). Analysis of randomized join-the-shortest-queue (jsq) schemes in large heterogeneous processor sharing systems. *IEEE Transactions on Control of Network Systems* **3(2),** 116–126.

[35] MUKHOPADHYAY, A., MAZUMDAR, R. R. AND GUILLEMIN, F. (2015). The power of randomized routing in heterogeneous loss systems. In *Teletraffic Congress (ITC 27), 2015 27th International*. pp. 125–133.

[36] MUKHOPADHYAYAY, A., KARTHIK, A., MAZUMDAR, R. R. AND GUILLEMIN, F. M. (September 2015). Mean field and propagation of chaos in multi-class heterogeneous loss models. *Performance Evaluation* **91,** 117–131.

[37] OELSCHLAGER, K. (1984). A martingale approach to the law of large numbers for weakly interacting stochastic processes. *Ann. Probab.* **12,** 458–479.

[38] ROBERT, P. (2003). Stochastic Modelling and Applied Probability Series. Springer-Verlag.

[39] ROSS, S. M. (2009). *Introduction to Probability Models*. Academic Press; 10th edition.

[40] RUDIN, W. (1987). *Real and complex analysis* third ed. McGraw-Hill Book Co., New York.

[41] SEVASTYANOV, B. A. (1957). An ergodic theorem for markov processes and its application to telephone systems with refusals. *Theory of Probability & Its Applications* **2,** 104–112.

[42] TURNER, S. R. E. (1996). Resource pooling in stochastic networks. *Ph.D. dissertation, University of Cambridge*.

[43] TURNER, S. R. E. (1998). The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences* **12,** 109–124.

[44] VASANTAM, T., MUKHOPADHYAY, A. AND MAZUMDAR, R. R. (2017). Mean-field analysis of loss models with mixed-Erlang distributions under power-of-d routing. In *2017 29th International Teletraffic Congress (ITC 29)*. vol. 1. pp. 250–258.

[45] VVEDENSKAYA, N. D., DOBRUSHIN, R. L. AND KARPELEVICH, F. I. (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Problems of Information Transmission* **32,** 20–34.

[46] XIE, Q., DONG, X., LU, Y. AND SRIKANT, R. (2015). Power of d choices for large-scale bin packing: A loss model. In *Proceedings of the 2015 ACM SIGMETRICS*. pp. 321–334.

[47] ZHANG, J. (2013). Fluid models of many-server queues with abandonment. *Queueing Systems* **73,** 147–193.